Avtomatska transkripcija in segmentacija za iskanje najbolj reprezentativnega dela v vokalnih ljudskih pesmih

Ciril Bohak

doktorska disertacija

predana

Fakulteti za računalništvo in informatiko

kot del izpolnjevanja pogojev za pridobitev naziva

doktor znanosti

s področja

računalništva in informatike

Ljubljana, 2016

IZJAVA

Izjavljam, da sem avtor dela in da slednje ne vsebuje materiala, ki bi ga kdorkoli predhodno že objavil ali oddal v obravnavo za pridobitev naziva na univerzi ali na drugem visokošolskem zavodu, razen v primerih, kjer so navedeni viri.

— Ciril Bohak —
junij 2016

Oddajo so odobrili

dr. Matija Marolt
docent računalništva in informatike
MENTOR IN ČLAN OCENJEVALNE KOMISIJE

dr. Danijel Skočaj
izredni profesor računalništva in informatike
PREDSEDNIK OCENJEVALNE KOMISIJE

dr. Andrej Košir
redni profesor elektrotehnike
ZUNANJI ČLAN OCENJEVALNE KOMISIJE
Univerza v Ljubljani, Fakulteta za elektrotehniko

PREDHODNA OBJAVA

Izjavljam, da so bili rezultati obravnavane raziskave predhodno objavljeni/sprejeti za objavo v recenzirani reviji ali javno predstavljeni v naslednjih primerih:

[1] Ciril Bohak, Matija Marolt. Probabilistic Segmentation of Folk Music Recordings. V *Mathematical Problems in Engineering*, Volume 2016 (2016), Article ID 8297987, 2016. Hindawi, http://dx.doi.org/10.1155/2016/8297987.

[2] Ciril Bohak, Matija Marolt. Transcription of folk music. V *Journal of the Audio Engineering Society*, [Poslano 13. 2. 2016], AES.

Potrjujem, da sem pridobil pisna dovoljenja vseh lastnikov avtorskih pravic, ki mi dovoljujejo vključitev zgoraj navedenega materiala v pričujočo disertacijo. Potrjujem, da zgoraj navedeni material opisuje rezultate raziskav, izvedenih v času mojega podiplomskega študija na Univerzi v Ljubljani.

To the principle of randomness and entropy in the universe, whose fault it is that you have a chance of reading this work, right here, right now.

FRQM TNIS WISTFYT FYTFXM CQINT TNM MFRTN MIXNT YQT SMMM QF FYE CFRTNCNFFR NTMRMST BNT FQR NS ITS WIFFMRMYT CQYSIWMR FXFIK TNFT WQT PNFTS NMRM PNFTS NQMM PNFTS NS QK IT MMRQQKM QQN FQM MMRQQKM QQN CYQP MMRQQKM QQN MMR NMFRDW QF MMRE NMMFK BMIYX PNQ MMR PFS NMW QNT TNMR NMS PNM FXXRMXFTM QF QNR QQE FYW SNFFMRIYX TNQNSFFWS QF CQYFIWMYT RMIXIQYS IWMQTQXIMS FYW MCQYQRMC WQCTRIYMS MMRE NNYMR FYW FQRFXMR MMRE NMRQ FYW CQPFRW MMRE CRMFTQR FYW WIMSTRQRMR QF CINYFTIQK MMRE CNIX FYW CMFSFYT MMRE QQNYX CQNCFM IS FQM MMRE MQTNMR FYW FFTNMR NQCMFNT CNIW NYMYQR FYW MCTQRMR MMRE TMFCNMR QF MQRFMS MMRE CQRRNCT CQNTMCIK MMRE SNCMRSTFR MMRE SNCRMMM TMFWMR MMRE SFIYT FYW SIYXMR IS TNM NISTQRE QF QNR SCMCIMS NMW TNMRM QK F MQTM QF WINST SNSCMYWMW IS F SNYBMFM

HNM MFRTN IS F MRE SMFNT STFXM IS F FST CQSMC FRMYF HNYC QF TNM RMRS QF BTQQW SCMMW BE FNT TNQSM XMYMRFMS FYW MMCMRQRS SQ TNFT IS XTQRE FYW TRNMCN TNME CQNPW BMCQMM TNM MQMMYTFRE MFSTMRS QF F FRFCTIQK QF F WQT HNYC QF TNM MYWMSSS CRNMTMS ISTMW BE TNM NYNFBITFKS QF QKM CQRYMR QF TNIS CMT QK TNM SCFRCMR WISTHXNSNFBTM NYNFBI TFKS QF SQMM QTNMR CQRYMR NQP FRMNMYT TNMR MISNYWMRSTFYWNYS NQP MFXMR TNME FRM TQ CMT QKM FYQTNMR NQP FMRMYT TNMR NFTRMWS QNR CQSTNRNYS QNR IMFXNMW SMTFMCQR TFKCM TNM WMNNSIQK TNFT PM NFM SQMM CRMMXMW CQSIMQK IS TNM NYMRSM FRM CNFTMYXMW BE TNIS CQYT QF CFM NXNT QNR CTFYMT IS F PQYMR SCMK IS TNM XRMFT MYMTQCNX CQSMC WFRC JK QNR QBSCNRME IS FNT TNIS FSTYMSS TNMRM IS YQ NNT TNFT NMC PNT CQMM FRQM MSMPNMRM TQ SFM NS FRQM QNRSMMS

HNM MFRTN IS TNM QKR PQRW CYQPK SQ FFR TQ NFRBQR NFM HNMRM IS YQPNMRM MSM FT MFST IS TNM YMFR FNTNRM TQ PNKN QNR SCMCIMS CQNW MIXRFTM ISIT QMS SMTTNM YQT QMT NCM IT QR YQT FQR TNM MQMMYT TNM MFRTN IS PNMRM PM MFCM QNR STFYW JT NFS BMMY SFIW TNFT FSTRQYQMR IS F NNMBNYX FYW CNFRFCMRBNNWYX MCMRIMYCM HNMRM IS CMRNFCS YQ BMTTMR WMMQYSTRFTIQK QF TNM FQTR QF NNMFK CQYCMITS TNFK TNIS WISTFYT IMFXM QF QNR TYKE PQRW PE PM IT NYWMRSCRMS QNR RMSCQYSIBINE TQ WMNFT MQRM CIWR PITN QKM FYQTNMR FYW TQ ERMSMRM FYW CNMRISN TNM CFM BNM WQT TNM QKR NQMM PMM MMR CYQPK

CFRT SFXFK CFM BNM WQT F ISIQK QF TNM NNMFK FNTNRM IS SCFCM

Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*

Ciril Bohak

Avtomatska transkripcija in segmentacija za iskanje najbolj reprezentativnega dela v vokalnih ljudskih pesmih

POVZETEK

Cilj glasbene segmentacije je razviti algoritme, ki bodo v zvočnem posnetku poiskali ponavljajoče vzorce glede na želeni aspekt (melodija, ritem, barva zvoka) in določili meje med posameznimi ponovitvami. Pri glasbeni transkripciji je cilj razviti algoritme, s katerimi lahko iz zvočnega posnetka pridobimo informacijo o prisotnosti višin tonov v posameznih časovnih okvirjih. Pri tem se lahko osredotočimo na monofonične posnetke ali na polifonične posnetke. Segmentacija in transkripcija predstavljata pomembna dela raziskovalnega področja pridobivanja informacij iz glasbe. Rezultati so uporabni za veliko realnih aplikacij; s segmentacijo glasbe delno določimo glasbeno strukturo pesmi, ugotovimo melodična ponavljanja v pesmih ali si pomagamo pri iskanju najbolj reprezentativnega dela pesmi; transkripcijo lahko uporabimo pri avotmatskem generiranju notnega zapisa, kot pomoč pri ročni transkripciji glasbe ali za iskanje podobnih melodij v glasbenih zbirkah.

V pričujoči doktorski disertaciji naslavljamo specifično problematiko tako segmentacije kot transkripcije zvočnih posnetkov, natančneje segmentacijo in transkripcijo zvočnih posnetkov ljudske glasbe. Že razvite metode na ljudski glasbi odpovedo zaradi njenih specifik, kot so slabi snemalni pogoji in amaterski izvajalci, zaradi česar prihaja do pojavov, kot so visoka stopnja šuma v posnetkih, netočno petje, drsenje višine tonov skozi pesem, neenakomeren tempo ipd. V uvodu podamo motivacijo za izpeljavo raziskav in podrobno opredelimo probleme in cilje.

Prvi del disertacije predstavi raziskave s področja glasbene segmentacije, kjer predstavimo metodo za segmentacijo ljudske glasbe, ki na zbirki ljudske glasbe deluje bolje od trenutno aktualnih segmentacijskih metod. Predstavljena segmentacijska metoda deluje na podlagi verjetnostnega modela za iskanje ponavljajočih melodičnih delov v

posnetku in določanje njihovih začetkov. Predstavljena metoda je bila ovrednotena na zbirki posnetkov ljudske glasbe različnih tipov: solo pesmi, dvo- in triglasne pesmi, zborovske pesmi, instrumentalne pesmi ter mešane pete in instrumentalne pesmi. Razvita metoda je ovrednotena tudi iz aspekta robustnosti, kjer smo preverjali odpornost razvite metode glede na degradacije.

V drugem delu disertacije predstavimo raziskave, povezane z glasbeno transkripcijo, kjer opišemo metodo za transkripcijo ljudskih pesmi. Metoda na podlagi segmentacije poišče reprezentativni del in ga s pomočjo vseh ponovitev znotraj pesmi transkribira. Metoda kot vhod prejme ocene že izračunanih osnovnih višin tonov in segmentacijo pesmi. Na podlagi segmentacije metoda medsebojno poravna vhodne višine tonov v časovni in frekvenčni domeni, odstrani lokalne nepravilnosti in združi transkripcijo vseh segmentov. V drugem koraku metoda izračuna note s pomočjo dvonivojskega verjetnostnega modela, temelječega na Skritih markovskih modelih z eksplicitno določenim trajanjem obiskov posameznih stanj, ki modelira oceno not, pavz in notnih prehodov. Predstavljena metoda je bila ovrednotena na zbirki večglasne ljudske glasbe, kjer vrača boljše rezultate od aktualnih trankripcijskih metod.

V zaključkih disertacije izpostavimo znanstvene prispevke ter podamo možnosti za nadaljnji razvoj in uporabo posamezne predstavljene metode.

Ključne besede: pridobivanje informacij iz glasbe, segmentacija glasbe, glasbena struktura, transkripcija, povzemanje glasbe

University *of Ljubljana*
Faculty *of Computer and Information Science*

Ciril Bohak
Finding the most representative part of vocal folksongs with transcription and segmentation

ABSTRACT

The goal of musical segmentation is to develop algorithms that will find similar patterns in audio signal according to desired aspect (melody, rhythm, timbre) and to define the boundaries between the repetitions. The goal of musical transcription is to develop algorithms that will extract pitches from the audio signal in every time frame either for monophonic or polyphonic music. Music segmentation and transcription represent two very important parts of music information retrieval research field. The results can be used in many real-life applications: with music segmentation we can define musical structure, melodic repetitions in music or we can use it in search for most representative part; transcription results can be used in automatic generation of scores, as a support in manual transcription process or in search of similar melodies in musical collections.

In the presented dissertation we are addressing specific problems of musical segmentation and transcription of audio recordings: segmentation and transcription of folk music audio recordings. Currently developed methods fail on folk music due to it's specifics, such as bad recording conditions and amateur performers, which are the reason for high level of noise in recordings, inaccurate singing, pitch drifting throughout the song etc. In introduction section we give the motivation for conducting the research and define the problems and goals of the thesis in the detail.

The first part of the dissertation presents the research from field of music segmentation, where we present a folk music segmentation method, that outperforms current state-of-the-art methods on a collection of folk music. The presented segmentation method bases on a probabilistic model for finding melodically repeating parts in recording and defining their beginnings. The method was evaluated on a folk music collection of different types: solo singing, two- and three-voiced singing, choir songs, instrumental

songs and mixed assembles. The developed method was also evaluated according to robustness aspect, where resistance to different degradations was tested and evaluated.

The second part of the dissertation addresses musical transcription, where we present a folk music transcription method. The method uses the segmentation results to find a representative part of a song and transcribes it with use of all the repetitions within the song. The method takes multiple fundamental frequencies estimations calculated with an existing method and song segmentation. With use of segmentation results the method aligns the multiple fundamental frequencies estimations in temporal and frequency domain, removes local inaccuracies and joins the transcriptions of all repeating parts. In next stage the method calculates notes using two-level probabilistic model based on explicit duration Hidden Markov models, used to model notes, rests and note transitions. The presented method was evaluated on collection of polyphonic folk music, where it returns better results of current state-of-the-art music transcription methods.

In the conclusions we highlight the scientific contributions of the thesis and give the directions for possible future improvements and extensions of the method.

Key words: music information retrieval, music segmentation, musical structure, transcription, audio fingerprinting

ZAHVALA

Doktorat znanosti predstavlja še zadnjo formalno stopničko izobrazbe, a zame vsekakor ne pomeni konca učenja in izobraževanja. Vsekakor predstavlja doktorat najvišjo stopničko, za katero je bilo potrebnega veliko truda, volje, pomoči in vzpodbude. Medtem, ko sem trud in voljo prispeval sam, pa se moram za pomoč in vzpodbudo zahvaliti kar nekaj ljudem. Na prvem mestu se želim zahvaliti svoji družini, ki me je vseskozi podpirala v mojih prizadevanjih in me vzpodbujala k temu, da to zadnjo stopničko čim prej osvojim. Prav tako so me na moji poti vseskozi podpirali tudi prijatelji, ki so mi s svojo podporo pomagali prebresti težja obdobja.

Velika zahvala gre mentorju Matiji Maroltu, ki je s svojim znanjem, karizmo izkušnjami in pomočjo zaslužen, da sem uspel doktorsko disertacijo uspešno zaključiti. Skupaj s sodelavci in študenti v Laboratoriju za računalniško grafiko in multimedije Fakultete za računalništvo in informatiko so mi skozi celotno obdobje zagotavljali sproščeno okolje z odličnim vzdušjem in karizmo. Hvala vam Matic, Alenka, Matevž, Saša, Pia, Žiga, Primož in Manca.

Zahvala gre tudi Glasbenonarodopisnemu inštitutu Znanstvenoraziskovalnega centra Slovenske akademije znanosti in umetnosti, ki so nam omogočili uporabo njihovih podatkov za pripravo glasbenih zbirk uporabljenih za testiranje in ovrednotenje metod razvitih v okviru doktorske disertacije.

*Nenazadnje se želim zahvaliti tudi svoji punci, ki je vzporedno z mano bila enako bitko a na drugi fronti. Skupno bojevanje in medsebojna podpora sta bila neizmeren vir energije, potrebne za uspešen zaključek. Hvala Vida :**

— Ciril Bohak, Ljubljana, junij 2016.

KAZALO

I

Uvod

The beginning is the most important part of the work.

– Plato, The Republic

Glasba - umetnost, katere izrazno sredstvo je zvok.

(vir. Slovar slovenskega knjižnega jezika)

Glasba človeka spremlja že od prazgodovine in predstavlja za marsikoga zelo pomemben del življenja. Nenazadnje se dandanes to odraža v dejstvu, da predstavlja glasbena industrija enega izmed pomembnih gospodarskih temeljev današnje družbe.

Skupaj s človekom se je skozi zgodovino spreminjala tudi glasba in načini njenega ustvarjanja. Od preprostih instrumentov v prazgodovini, starih preko 50.000 let [Turk and Kavur, 1997], so ljudje skozi čas izumljali nove in nove glasbene instrumente, ki so dandanes zgrajeni tudi z uporabo najsodobnejših tehnologij. Vseskozi je za enega izmed glavnih načinov ustvarjanja glasbe vsekakor veljal, in še vedno velja, človeški glas. Nenazadnje predstavlja človeški glas tudi temeljni način komunikacije z ostalimi kot tudi način izražanja stanja posameznika. Dolgo časa je za edini način ohranjanja glasbe veljalo neposredno izročilo, saj ljudje do antične Grčije niso izumili nekega formalnega načina za zapis glasbe. V antični Grčiji so zapis glasbe definirali s spreminjanjem melodije s posebnimi oznakami nad glasbenimi besedili, kar pa je zgolj nakazovalo spremembo višine tonov skozi čas, ne pa tudi njene osnovne izhodiščne vrednosti. Notna oblika zapisa glasbe, podobna, kot jo poznamo danes, se je pojavila šele v poznem srednjem veku.

Kljub temu da je od srednjega veka dalje obstajala možnost formalnega zapisa glasbe, pa le-tega niso poznali preprosti ljudje, ampak zgolj ustrezno izobražena manjšina. Vseeno pa je bila glasba razširjena tudi med ljudstvom in se je ohranjala enako kot do tedaj - z ustnim izročilom. Takšno glasbo imenujemo ljudska glasba. Uporaba ustnega izročila je tudi eden izmed glavnih razlogov, zakaj lahko danes naletimo na toliko različnih izvedenk istih pesmi, ki krožijo med ljudmi na različnih območjih.

Šele v poznem 19. stoletju so pričeli muzikologi zbirati in ohranjati ljudsko glasbo. Sistematični popis ljudske glasbe se je izvajal z obiskom posameznikov in skupin na različnih območjih z namenom dokumentiranja pesmi v pisni, kasneje pa tudi zvočni obliki skupaj s podatki o lokaciji, izvajalcih in še množico ostalih informacij. Z napredkom tehnologije - prihodom računalnikov - se je moderniziral tudi postopek zbiranja

in obdelave arhiviranih pesmi z namenom izvedbe različnih raziskav. Prvi korak je predstavljala pretvorba obstoječih arhivov v digitalno obliko za boljšo dostopnost in lažji vpogled.

S tem se je ustvarilo tudi novo raziskovalno področje - pridobivanje informacij iz glasbe (angl. music information retrieval - MIR) [Taque-Sutcliffe et al., 1993]. MIR predstavlja izredno raznoliko multidisciplinarno raziskovalno področje, ki pokriva tako raziskave, povezane z glasbo v vseh oblikah, kot tudi raziskave na z glasbo povezanih podatkih. V zadnjih letih je raziskovalcem na področju MIR na voljo vedno več javno dostopnih odprtih podatkovnih zbirk ljudske glasbe, ki so bile vzpostavljene v okviru različnih projektov za ohranjanje kulturne dediščine. Nekaj primerov takšnih zbirk predstavljajo: *Essen Folksong Database* [Schaffrath, 1995], ki vsebuje preko 20.000 večinoma evropskih ljudskih pesmi; zbirka *Finish Folk Tunes* [Toiviainen and Eerola, 2004], ki vsebuje preko 9.000 finskih ljudskih pesmi; zbirka *American Folk Song Collection* [Center, 2004]; zbirka *Australian Folk Songs* [of Australia, 1994]; nenazadnje med takšne zbirke sodi tudi zbirka slovenskih ljudskih pesmi *Etnomuza*[1] [Strle and Marolt, 2007], ki vsebuje slovensko ljudsko glasbo iz Slovenije in zamejskih področij. Nekoliko novejša zbirka anotiranih materialov *The Meertens Tune Collections* je predstavljena v [van Kranenburg et al., 2014]. Posamezne zbirke vsebujejo različne vsebine; nekatere zbirke so izredno dobro strukturirane, medtem ko druge sestojijo iz zelo slabo strukturiranih podatkov v najrazličnejših oblikah.

Medtem ko je veliko raziskav usmerjenih v razvoj algoritmov za pridobivanje informacij iz komercialnih glasbenih posnetkov, kjer (odvisno od namena) lahko delujejo zelo dobro, te iste metode v večini primerov delno ali povsem odpovedo pri posnetkih ljudske glasbe. Razloge lahko poiščemo v specifikah ljudske glasbe, predvsem v slabih snemalnih pogojih in amaterskih izvajalcih, zaradi česar prihaja do pojavov, kot so visoka stopnja šuma v posnetkih, netočno petje, drsenje višine tonov skozi pesem, neenakomeren tempo ipd. Veliko specifik ljudske glasbe je predstavljenih tudi v rezultatih projekta *WITCHCRAFT* [Kranenburg et al., 2007]. Ta dejstva so tudi glavni vzrok, da se v zadnjem času raziskovalci področja MIR vse bolj posvečajo tudi domeni ljudske glasbe in razvijajo pristope in metode, namenjene posebej na ljudski glasbi.

[1] Zbirka *Etnomuza* je dostopna preko rezultatov projektov *Klik v domovino* in *EtnoFletno*

1.1 Motivacija

Posamezniku je danes na razpolago vedno več organiziranih in urejenih glasbenih zbirk, v katerih je vsebina predstavljena strukturirano in večinoma na intuitiven način. Med takšne zbirke lahko prištevamo tudi arhive ljudske glasbe, med katerimi nekateri vsebujejo več tisoč posnetkov. Za takšne zbirke velja, da je njihovo ročno urejanje in strukturiranje časovno izredno potratno. Z uporabo sodobnih tehnologij pa je mogoče določene naloge avtomatizirati in s tem delo pohitriti, poenostaviti, hkrati pa zmanjšati možnost vnosa človeške napake. Kljub temu pa še vedno ostaja veliko nalog, ki bi jih želeli avtomatizirati, a za to še niso bili razviti dovolj dobri postopki. Z nekaj takšnimi problemskimi domenami se soočamo v tem delu.

Pri urejanju arhivskih zapisov muzikologi iz daljšega zvočnega posnetka ročno izrežejo del posnetka, ki predstavlja posamezno pesem. Vsako dobljeno pesem nato podrobno analizirajo. Analiza vključuje tako zapis besedila, transkripcijo melodije, število ponovitev kitic v pesmi, število in tip pevcev ipd. Muzikologi vse omenjene naloge še vedno večinoma rešujejo ročno. V okviru raziskovalnega dela doktorske disertacije smo naslovili nekaj od predstavljenih problemov in zanje podali avtomatizirane rešitve.

Prvi problem, ki ga naslavljamo, je iskanje ponavljajočih delov znotraj zvočnega posnetka posamezne pesmi. Ta problem sodi v kategorijo ugotavljanja glasbene strukture pesmi. V našem primeru, ko se ukvarjamo z ljudsko glasbo, to predstavlja ponavljajoče kitice. Cilj avtomatskega pristopa je pridobiti čimbolj točne začetke posameznih kitic znotraj pesmi in posledično tudi njihovo število. Sama struktura ljudskih pesmi je sicer dokaj enostavna - melodično ponavljajoči deli - a izkaže se, da zaradi narave materialov - ljudska glasba - to še zdaleč ni enostavno rešljiv problem, kar se pokaže v tem, da obstoječi pristopi, razviti za popularno, klasično ali rock glasbo, odpovedo. V okviru disertacije predstavimo izvirni pristop k segmentaciji ljudske glasbe, ki je robusten in ki se kosa z najboljšimi na področju in jih v marsikaterem primeru tudi prekosi.

Drugi problem, naslovljen v tem delu, predstavlja avtomatsko transkripcijo ljudske glasbe. Problem se izkaže še za posebej težavnega v primerih večglasnega petja, kar je pri ljudski glasbi pogosto. Cilj je izdelati čim boljšo notno predstavitev melodije, ki je odpeta ali odigrana v posamezni pesmi. Podobno kot pri prvem problemu se tudi tukaj izkaže, da obstoječi pristopi ne vračajo zadovoljivih rezultatov, ponovno zaradi

same narave materialov. Drugi del disertacije smo posvetili reševanju tega problema in razvili izvirni pristop k transkripciji ljudske glasbe, ki presega rezultate najboljših pristopov na področju.

Tretji problem, s katerim se ukvarjamo v okviru disertacije, naslavlja iskanje najbolj reprezentativnega dela znotraj izvedbe posamezne ljudske pesmi in njene transkripcije. Najbolj reprezentativni del naj bi predstavljal najbolj tipično izvedbo (ponavljajočega) dela pesmi in njeno transkripcijo.

V okviru disertacije pojasnimo tudi, kako se predstavljeni problemi med seboj prepletajo in kako lahko rešitvi prvih dveh problemov uporabimo pri reševanju tretjega problema.

1.2 Cilji in struktura disertacije

V nadaljevanju dela je predstavljen strukturiran pregled raziskovalnega področja s poudarki na delih, posebej pomembnih za raziskave, povezane s problematikami, ki jih naslavlja disertacija. Sledita dva dela, ki predstavita izvirne znanstvene prispevke za rešitev predstavljenih problemov.

V prvem je podrobneje predstavljena problematika segmenetacije glasbe. Predstavljen je izvirni pristop za avtomatsko segmentacijo ljudske glasbe. Predstavljen pristop je ovrednoten na zbirki ljudskih pesmi, na kateri smo ovrednotili tudi nekatere trenutno najboljše pristope in rezultate primerjali z rezultati predstavljenega pristopa za segmentacijo ljudske glasbe. Prav tako je pristop ovrednoten s stališča robustnosti.

Izvirni znanstveni prispevek:

- Razvit robusten algoritem za segmentacijo ljudske glasbe, ki naslavlja probleme ljudskih pesmi.

Drugi del predstavi problematiko glasbene transkripcije in izpostavi probleme, povezane z ljudsko glasbo. Predstavljen je izvirni pristop k transkripciji ljudske glasbe. Omenjeni pristop je v nadaljevanju tudi ovrednoten na zbirki večglasnih ljudskih pesmi, na kateri smo ovrednotili tudi nekaj trenutno najboljših pristopov in njihove rezultate

primerjali z rezultati predstavljenega pristopa za transkripcijo ljudske glasbe. Prav tako je pristop ovrednoten s stališča robustnosti.

Izvirni znanstveni prispevek:

- Robusten algoritem za transkripcijo vokalnih ljudskih pesmi, odporen na šume in prekinitve v posnetkih, ki upošteva netočno petje izvajalcev.

Oba dela skupaj predstavita tudi pristop za iskanje in transkripcijo najbolj reprezentativnega dela pesmi. S tem pridobimo tako del zvočnega posnetka, ki je najbolj reprezentativen za celoten posnetek, kot tudi njegovo transkripcijo.

Izvirni znanstveni prispevek:

- Na podlagi prejšnjih prispevkov razvit pristop za iskanje najbolj reprezentativnega dela v zvočni domeni ter njegove transkripcije.

V zaključnem poglavju predstavimo zaključke obeh delov, predstavimo izvirne znanstvene prispevke disertacije in navedemo izhodišča za morebitne nadaljnje raziskave.

2

Pregled področja

To know that we know what we know, and to know that we do not know what we do not know, that is true knowledge.

– Nicolaus Copernicus

Digitalna revolucija širi svoj vpliv na številna področja. Posledično izredno narašča tudi t.i. digitalizacija - pretvarjanje oz. shranjevanje podatkov v digitalni obliki. Takšni podatki pa se hranijo v digitalnih zbirkah podatkov. Digitalizacija vse bolj zajema tudi arhivske zbirke podatkov, med katere uvrščamo tudi glasbene arhive. Sam proces digitalizacije arhivov in zbirk je zaradi ohranjanja kulturne dediščine tudi podprt s finančnimi sredstvi tako posameznih držav kot tudi na primer s strani Evropske unije. Prav na področju glasbe se večina novih zapisov že zajema in shranjuje v digitalni obliki, zaradi česar se že obstoječe zbirke dopolnjujejo in tudi povečujejo.

Javno dostopne zbirke glasbe ne služijo zgolj primarnemu namenu - ohranjanju dediščine za zanamce - ampak vzpodbujajo tudi sekundarne uporabe in aktivnosti, kamor lahko uvrščamo tudi znanstvene raziskave na glasbenih zbirkah. Takšne raziskave služijo različnim namenom kot so na primer: lažje in hitrejše iskanje po zbirkah; iskanje povezav med posameznimi zapisi v zbirki na podlagi različnih podobnostnih mer in s tem ugotavljanje izvora in povezanosti posameznih zapisov; urejanje podatkov na podlagi določenih lastnosti ipd. V takšne raziskave so lahko vključeni znanstveniki različnih znanstvenih disciplin: muzikologi, etnomuzikologi, sociologi, računalničarji in ostali. To interdisciplinarno raziskovalno področje imenujemo področje pridobivanja informacij iz glasbe.

2.1 *Pridobivanje informacij iz glasbe*

Interdisciplinarno področje pridobivanja informacij iz glasbe [Taque-Sutcliffe et al., 1993] (MIR), ki združuje vede muzikologije, bibliotekarstva in urejanja zbirk, pridobivanja informacij in računalništva, je področje, kamor uvrščamo pričujočo disertacijo. Področje se ukvarja z različnimi načini obdelave glasbe v digitalni obliki in med drugim zajema tudi:

- metode za razpoznavanje glasbenih vzorcev, izračun značilnic, računanje podobnosti, razvrščanje v razrede, gručenje;
- metode za prepoznavanje glasbe in izvajalcev, sledenje besedilu in notnemu zapisu, avtomatsko spremljavo, iskanje na podlagi glasbenih poizvedb;
- prepoznavanje sestavov in glasbenih instrumentov v zvočnih posnetkih, pridobivanje informacije o ritmu, tempu, melodiji, harmoniji, barvi zvoka …

Raziskave s področja MIR večinoma naslavljajo popularno in klasično glasbo. Vzrok temu je predvsem dejstvo, da so bile, in še vedno so, zbirke popularne in klasične glasbe večje in bolj dostopne. Prav tako so raziskave na popularni in klasični glasbi komercialno bolj zanimive in s tem pritegnejo tudi glasbena podjetja in založbe, ki takšne raziskave financirajo. Stvari se v zadnjih letih spreminjajo predvsem zaradi večje dostopnosti zbirk ostale glasbe, na primer ljudske glasbe. Prav dostopnost zbirk je eden izmed razlogov za porast raziskav z ljudsko glasbo.

Informacije, pridobljene iz glasbe, lahko uporabimo, kadar želimo na podlagi določenih lastnosti glasbe sprejemati odločitve. Na podlagi informacij lahko iz neke glasbene zbirke izločimo tiste pesmi, katerim je skupna določena lastnost: so na primer iste zvrsti ali pa imajo podoben tempo. Na podlagi takšnih lastnosti lahko sestavimo prilagojene sezname za predvajanje, ki ustrezajo želenim kriterijem poslušalca. Prav tako lahko na podlagi informacij, pridobljenih iz glasbe, ki jo nek uporabnik posluša, sklepamo na njegov glasbeni okus in mu priporočimo glasbo s podobnimi lastnostmi, ki je še ne pozna. Na podlagi pridobljenih informacij lahko sklepamo tudi na to, katera glasba je vplivala na katerega avtorja pri sestavljanju njegove glasbe. Takšnih in podobnih scenarijev si lahko zamislimo še mnogo. Sistemi, ki počno predstavljene stvari, so v uporabi že nekaj let in dobro služijo uporabnikom.

Zaradi interdisciplinarnosti področja MIR lahko z rezultati metod s tega področja pomagamo tako raziskovalcem s področja glasbe kot tudi glasbenim ustvarjalcem in nenazadnje glasbeni industriji in poslušalcem. Kar nekaj pristopov lahko tako srečamo tudi v vsakodnevni uporabi. Takšna je na primer storitev Shazam[1], ki uporabniku omogoča, da na podlagi kratkega posnetka, posnetega v vsakdanjem okolju, poizve, katera pesem se predvaja v njegovi okolici. Drug primer takšne aplikacije je odkrivanje kršitev pri uporabi oz. objavi licenčne glasbe na spletu, kar že nekaj časa s pridom uporabljajo spletni portali z video in zvočnimi vsebinami, kot so YouTube[2], Yousician[3], Flowkey[4], Spotify[5] in drugi.

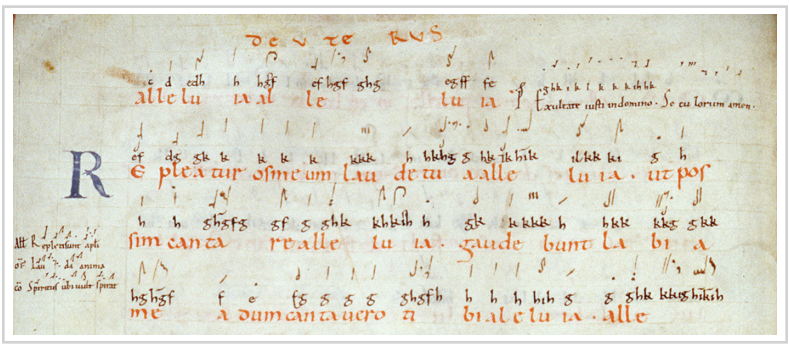
Zelo pomembno je tudi, kako dobro delujejo posamezni razviti pristopi zato so v

[1] http://www.shazam.com/
[2] https://www.youtube.com
[3] https://www.yousician.com
[4] https://www.flowkey.com
[5] https://www.spotify.com

ta namen začeli prirejati tekmovanja (angl. Music Information Retrieval Evaluation eXchange - MIREX) [Downie, 2008] z namenom primerjave posameznih pristopov za reševanje različnih problemov s področja MIR. Tekmovanje je namenjeno primerjavi in evalvaciji posameznih pristopov na istih zbirkah podatkov, saj lahko le tako ugotovimo, kateri pristopi delujejo bolje od ostalih.

2.2 *Glasbeni zapis*

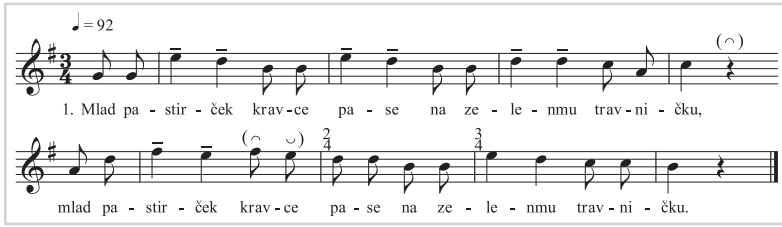
Človek ustvarja in posluša glasbo že od pradavnine. Zapis glasbe pa je doživljal podoben razvoj kot zapis jezika v obliki pisave. Prvi glasbeni zapisi segajo tako v obdobje 2.000 let pred našim štetjem. To so zapisi na tablicah, ki povzemajo osnovne glasbene lastnosti. Skozi obdobja antike in bizantinskega cesarstva se je glasbeni zapis precej spreminjal in je v srednjem veku pridobil danes skoraj vsem znano obliko notnega črtovja. Primer zapisa glasbe iz bizantinskega obdobja je prikazan na sliki 2.1.



Slika 2.1: Primer glasbenega zapisa iz bizantinskega obdobja[6].

Tudi notno črtovje je doživljalo spremembe in se je skozi čas iz začetne oblike s štirimi črtami preoblikovalo v notno črtovje s petimi črtami. Do zapisa, kot ga uporabljamo danes, se je zgodilo še precej sprememb. Spremenile so se oblike not, uveljavili so se dodatni simboli, kot so na primer simboli za označevanje takta in pavz. Primer takšnega zapisa je podan na sliki 2.2. Sam notni zapis se še danes spreminja in izpopolnjuje. Nekateri umetniki vpeljujejo lastne oznake in simbole, s katerimi želijo zapisati lastnosti, neopisljive z obstoječim naborom simbolov.

[6] vir: F-MOf H159: Montpellier, Bibliothèque Interuniversitaire, Section Médecin, Ms. H159, fol. 25v



Slika 2.2: Primer sodobnega glasbenega zapisa v obliki notnega črtovja[7].

Takšen zapis glasbe ponazarja enega izmed bolj pogostih načinov zapisa. K samemu zapisu glasbe večinoma pripišemo tudi pripadajoče besedilo, v kolikor ga glasba vsebuje. Zapis s pomočjo notnega črtovja predstavlja zgoščen način zapisa glasbe in omogoča enostavno ponovitev izvedbe. Kljub vsemu pa posameznemu izvajalcu še vedno dopušča lastno interpretacijo zapisa in s tem dodajanja lastnega pečata k izvedbi glasbe. Večina zapisov glasbe v simbolični obliki pa ne vsebuje dovolj informacij za enolično ponovitev izvedbe. Posledično se lahko izvedbi dveh izvajalcev med seboj precej razlikujeta, kljub težnji po čim bolj enakem izvajanju. Izjemo v tem primeru predstavljajo mehanični stroji (t.i. glasbeni avtomati), ki so sposobni na podlagi simboličnega zapisa glasbe, večinoma zapisanega v obliki notnega traku, enakovredno reproducirati glasbo tudi večkrat. Primer predstavitve v obliki notnega traku je prikazan na sliki 2.3.

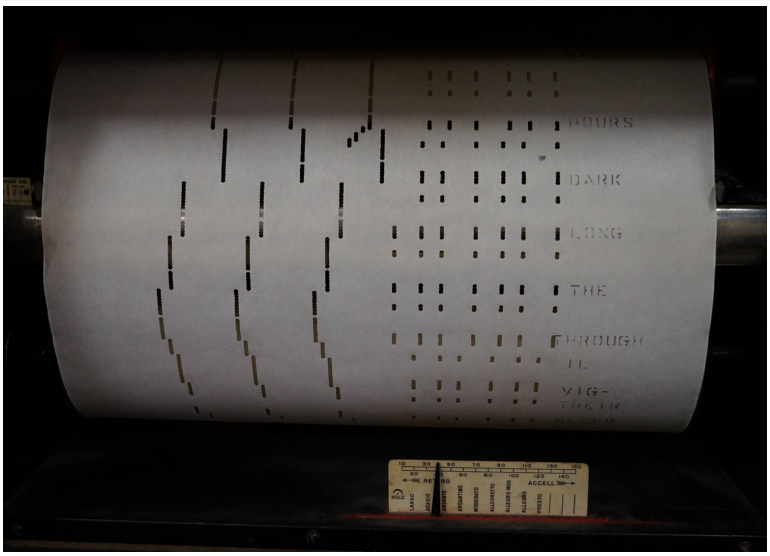
Ravno zaradi enostavne ponovne reprodukcije glasbe so se sčasoma pojavili tudi drugačni zapisi glasbe, na primer v obliki zvočnih posnetkov. Takšni posnetki so bili sprva posneti v analogni tehniki in so omogočali preprosto reprodukcijo posnete glasbe z uporabo ustrezne naprave (npr. fonograf ali gramofon). Z razmahom digitalne tehnike so se pojavili tudi digitalni načini zapisa glasbe tako v simbolični obliki kot v zvočni obliki. Primer digitalnega simboličnega zapisa predstavlja na primer zapis v formatu MIDI[9]. Primer digitalnega zvočnega zapisa pa predstavljajo zvočni posnetki v formatu WAVE[10].

[7] vir: glasbeni arhiv Etnomuza

[8] vir: Remembering the pianola - Lynn Buckler Walsh, https://lynnwalsh.wordpress.com/tag/pianola-roll/ - dostopano 9. junij 2016.

[9] angl. Musical Instrument Digital Interface - opis formata je dostopen na spletnem naslovu: http://www.midi.org/ - dostopano 9. junij 2016.

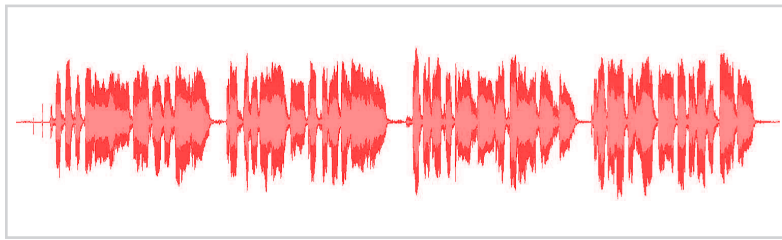
[10] angl. Waveform Audio File Format - opis formata je dostopen na spletnem naslovu: http://www-mmsp.ece.mcgill.ca/Documents/AudioFormats/WAVE/WAVE.html - dostopano 9. junij 2016.



Slika 2.3: Primer simbolične predstavitve glasbe v obliki notnega traku[8].

Zvočni glasbeni posnetek nam poleg osnovnih glasbenih lastnosti omogoča tudi hranjenje dejanskih akustičnih lastnosti snemalnega prostora, lastnosti snemalne opreme in lastnosti uporabljenih glasbenih instrumentov. Velja tudi, da sta kvaliteta in lastnosti zvočnega posnetka odvisna od snemalne opreme in tudi od snemalnega okolja. Medtem ko je pretvorba zvočnega signala iz simbolične oblike v zvočni posnetek dokaj enostaven postopek, kjer moramo zgolj definirati, kako izvesti posamezen dogodek, opisan z določenim simbolom, pa ne velja enako za obratni postopek. Pretvorbo zvočnega posnetka v simbolični zapis imenujemo transkripcija in velja za veliko težji problem.

Zvočni posnetek je v računalniku v osnovni obliki shranjen kot valovna krivulja, ki predstavlja spremembe pritiska v snovi, skozi katero se zvok širi - v večini primerov to pomeni spremembo pritiska v zraku. Primer takšne krivulje je prikazan na sliki 2.4 in prikazuje nekomprimirano obliko zapisa zvočnega posnetka v računalniku.



Slika 2.4: Primer valovne krivulje zvočnega signala.

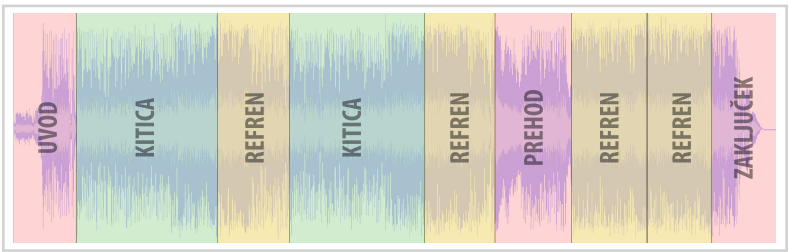
2.3 Glasbena struktura

Ravno zaradi svoje strukture - načina ponavljanja posameznih glasbenih vzorcev na različnih nivojih skozi čas - je glasba še posebej zanimiva za ljudi [Lerdahl and Jackendoff, 1983]. Najlaže si je takšno strukturo predstavljati za glasbo v simbolični obliki, kjer iščemo enaka (ali podobna) zaporedja skozi časovno dimenzijo: ponavljanje zaporedja not, akordov, intervalov in podobno. Glasbeno strukturo lahko definiramo na različnih nivojih [Deliége et al., 1996]. Avtor analizira razlike in povezave med površinsko strukturo in globoko strukturo glasbe. Skozi čas se lahko v glasbi ponavljajo krajši odseki (npr. nekaj not ali nekaj sekund zvočnega posnetka), lahko pa se ponavljajo tudi daljši odseki (npr. daljša notna zaporedja - motivi, kitice ali nekaj deset sekund dolgi odseki zvočnega posnetka).

Takšna struktura je značilna tako za klasično glasbo kot tudi za popularno glasbo. V primeru popularne glasbe takšne dele večinoma imenujemo *uvod*, *kitica*, *refren*, *prehod* in *zaključek*. Primer takšne glasbene strukture je prikazan na sliki 2.5. Izjemoma lahko popularna glasba vsebuje tudi drugačne odseke, za ljudsko glasbo pa praviloma velja, da ima še enostavnejšo strukturo, kjer se večinoma ponavljajo zgolj posamezne kitice, ki pa imajo enako melodijo.

Zakaj človek v glasbi uživa, še vedno ni povsem jasno. Zagotovo pa igra pri tem pomembno vlogo tudi glasbena struktura, saj so ravno vzorci tisti, ki človeka pritegnejo:

> " *Most adults have some childlike fascination for making and arranging larger structures out of smaller ones. One kind of musical understanding involves building large mental structures out of smaller, musical parts. Perhaps the drive to build those mental music structures is the same one that makes*



Slika 2.5: Primer tipične strukture pesmi popularne glasbe.

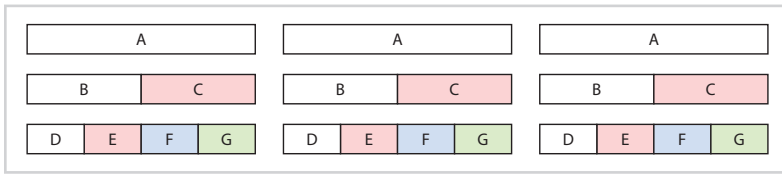
> *us try to understand the world. (Or perhaps that drive is just an accidental mutant variant of it; evolution often copies needless extra stuff, and minds so new as ours must contain a lot of that.)* "
>
> - Marvin Minsky, *Music, Mind, and Meaning [Minsky, 1981]*

Disertacija naslavlja podpodročje analize glasbene strukture - segmentacijo glasbe, ki je podrobneje predstavljeno v nadaljevanju, izvirni prispevki pa so predstavljeni v prvem delu disertacije.

2.3.1 Segmentacija glasbe

Večina obstoječih pristopov za segmentacijo zvočnih posnetkov naslavlja problem segmentacije popularne in klasične glasbe. Pregled aktualnih obstoječih metod za segmentacijo podajo avtorji v [Paulus et al., 2010], kjer so predstavljeni rezultati obstoječih pristopov segmentacije in iskanja struktur. Večina pristopov za segmentacijo glasbe temelji na uporabi izračunanih zvočnih značilnic, uporabljenih pri izračunu mer podobnosti med posameznimi deli posnetka, kar je vhod segmentacijskega algoritma.

Kot se izkaže, je lahko evalvacija segmentacije precej zapletena, saj lahko isto pesem segmentiramo na več nivojih. Takšen primer je prikazan na sliki 2.6. Poraja se namreč vprašanje, na katerem nivoju izvesti evalvacijo, nekaj predlogov podajajo avtorji v delu [McFee et al., 2015].



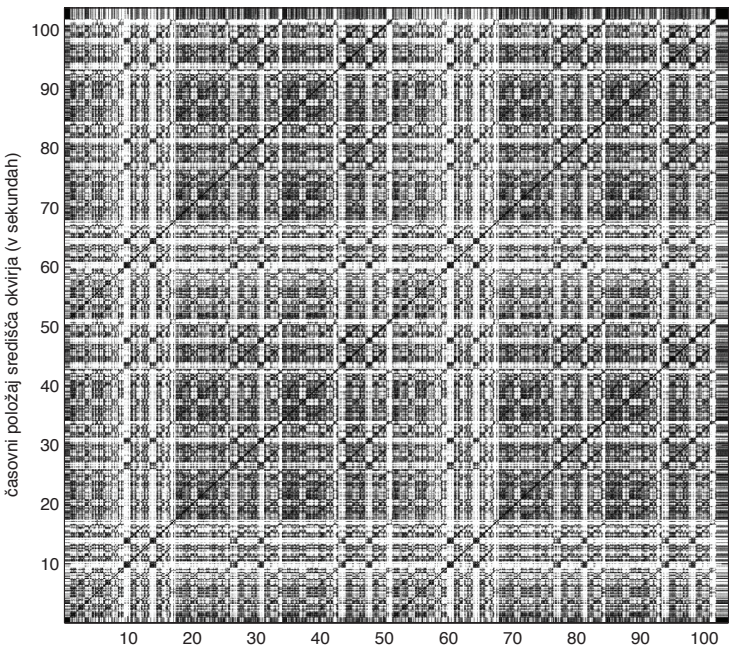
Slika 2.6: Primer glasbene strukture na različnih nivojih.

2.3.2 *Uporaba samopodobnosti*

Pri iskanju glasbene strukture so se raziskovalci sprva močno zanašali na iskanje ponavljajočih delov v različnih podobnostnih predstavitvah. Ena takšnih predstavitev je t.i. samopodobnostna matrika. Takšne matrike prikazujejo podobnost posameznih delov znotraj pesmi glede na lastnosti, na katerih so izračunane. Primer takšne matrike je prikazan na sliki 2.7, kjer je na posamezni osi časovna komponenta, intenziteta v matriki pa predstavlja podobnost med ustreznimi elementi.

Uporabo samopodobnostnih matrik je na področje MIR vpeljal Foote v delih [Foote, 1999, 2000]. V delu [Foote, 2000] avtor samopodobnosti znotraj pesmi išče na osnovi spektralnih lastnosti signala, podrobneje med seboj primerja spektralne lastnosti krajših oken, izračunanih s pomočjo hitre Fourierove transformacije (angl. fast fourier transform - FFT). Nad tako pridobljeno samopodobnostno matriko v nadaljnjih korakih izvede filtriranje, s katerim določi meje med posameznimi deli posnetka z enakimi lastnostmi. To predstavlja osnovo za številne nadaljnje pristope, ki pristop razširijo in uporabijo za segmentacijo razlničnih zvrsti glasbe. Eden takšnih pristopov je predstavljen v delu [Foote, 1999], kjer avtor za izračun samopodobnosti uporabi informacijo o barvi zvoka v določenem delu signala. Pristop uporabi za segmentacijo zvočnega posnetka pesmi v ponavljajoče dele glede na barvo zvoka.

Podoben pristop predstavita Cooper in Foote v delih [Cooper and Foote, 2002, 2003], kjer avtorja s pomočjo podobnosti v samopodobnostni predstavitvi poiščeta najbolj reprezentativen del zvočnega posnetka. Podobno metodo za avtomatsko generiranje glasbenih zvočnih povzetkov (angl. music audio summary) predstavijo avtorji v [Peeters et al., 2002]. Metoda iz glasbenega posnetka, na podlagi sampodobnostne predstavitve, izračunane na lastnostnih barve zvoka, izlušči glasbeni zvočni povzetek, ki zajema najbolj pogosto ponovljene in značilne dele celotnega zvočnega posnetka.



Slika 2.7: Primer podobnostne matrike.

Uporabo drugačnih značilnic v podobnem pristopu demonstrirata avtorja [Bartsch and Wakefield, 2001], ki v svojem delu za izračun samopodobnostne predstavitve uporabita kromatične značilnosti zvočnega signala. Temu primerno prilagodita tudi končno segmentacijsko metodo. Pristop, kjer je za izračun samopodobnosti uporabljena tako informacija o barvi zvoka kot tudi o kromatične značilnostih zvočnega signala, je predstavljen v [Eronen, 2007].

Pristop, kjer je samopodobnostna predstavitev izračunana za drugačne lastnosti zvočnih posnetkov, natančneje za ritem, predstavijo avtorji v delu [Foote et al., 2002]. Prav tako predstavijo nov koncept predstavitve tempa v posnetku, poimenovano ritmogram. Ogrodje za segmentacijo glasbenih zvočnih posnetkov z uporabo dekompozicije samopodobnostne predstavitve je predstavljeno v delu [Foote and Cooper, 2003]. Avtorja

podata primer uporabe ogrodja za iskanje najbolj reprezentativnega dela posnetka.

Jensen v svojem delu [Jensen, 2005] predstavi pristop, kjer samopodobnostne strukture izračuna na podlagi ritmičnih lastnosti glasbe. V kasnejšem delu pristop nadgradi tako, da združi lastnosti treh različnih domen [Jensen, 2007]: barve zvoka, ritma in harmonije. Na podlagi pridoblejnih samopodobnostnih struktur segmentira popularno glasbo na dele *uvod*, *refren*, *kitica* in *zaključek*. Avtor medsebojno primerja delovanje predstavljene metode z uporabo različnih značilnic in glede na ročno segmentacijo na zbirki 48-ih pesmi. Pri tem uporabi standardne mere za evalvacijo in pride do sklepa, da je segmentacija najboljša z upoštevanjem značilnic, ki modelirajo barvo zvoka.

Nekoliko drugačna predstavitev samopodobnosti za segmentacijo, ki vsebuje tudi prikaz časovnega zamika, je predstavljena v [Goto, 2003]. Metoda upošteva tudi morebitne spremembe v modulaciji. Avtor predstavi možne aplikacije predstavljene metode v delu [Goto, 2006] na zvočnih posnetkih popularne glasbe. Predstavljena metoda za razliko od večine ostalih dobro zazna tako začetke kot konce ponavljajočih delov - refrena. Avtorji so metodo ovrednotili na zbirki 100 posnetkov. V 80 primerih je metoda pravilno zaznala refren in njegove ponovitve. Prav tako je v delu predstavljena integracija metode v glasbeni predvajalnik.

Nov nabor značilnic za izračun samopodobnostne predstavitve, ki poleg značilnic za barvo zvoka zajema še značilnice za modeliranje višine tonov, je Peeters [Peeters, 2007] uporabil za nadaljnjo obdelavo samopodobnostne predstavtitve z namenom ojačanja izraženih ponovitev in zniževanja vpliva šuma. V tako izboljšani samopodobnostni matriki detektiramo posamezne ponavljajoče segmente z uporabo pristopa največjega verjetja (angl. maximum likelihood approach). Na manjši zbirki 11 popularnih pesmi se metoda odreže s 54,8 % uspešnostjo.

2.3.3 Uporaba matrične dekompozicije

Nedavno se je med metodami za odkrivanje strukture v glasbenih posnetkih precej uveljavila metoda nenegativne matrične faktorizacije (angl. non-negative matrix factorization - NMF). Ta pristop za odkrivanje strukture v popularni glasbi predstavita avtorja v delu [Kaiser and Sikora, 2010]. Avtorja uporabita NMF za dekompozicijo samopodobnostne matrike, izračunane na podlagi akustičnih lastnosti posnetka. Po-

kažeta tudi, da lahko z ustrezno faktorizacijo uspešno modeliramo posamezne dele strukture. Nadalje predstavita algoritem gručenja, ki pojasni celotno strukturo glasbenega posnetka, kar podkrepita z obetavnimi rezultati preliminarne evalvacije. Metoda na zbirki popularne glasbe skupine The Beatles doseže vrednost mere F1 0,62.

Pristop [Weiss and Bello, 2010] prav tako uporablja NFM, vendar za iskanje ponavljajočih vzorcev v samopodobnostni matriki, izračunani na z ritmom sinhroniziranih kromatičnih zvočnih značilnicah. Avtorja predstavita nenadzorovan pristop z uporabo konvolucijske nenegativne matrične faktorizacije, ki z upoštevanjem razpršenosti avtomatsko ugotovi število ponavljajočih vzorcev in njihovo dolžino. Segmentacijske metode te parametre navadno pridobijo drugače ali pa so to celo parametri same implementacije. Uspešnost metode na zbirki popularne glasbe skupine The Beatles doseže vrednost mere F1 0,6.

Avtorja [Nieto and Jehan, 2013] predstavita uporabo prilagojene izvedenke NMF - konveksno nenegativno matrično faktorizacijo (angl. convex non-negative matrix factorisation - CNMF) za avtomatsko identifikacijo glasbene strukture v popularni zahodni glasbi. Pristop v procesu faktorizacije uporablja konveksne omejitve, s čimer matriko razstavi na centroide, ki jih lahko interpretiramo kot posamezne dele pesmi. Metoda na bazi popularne glasbe skupine The Beatles doseže vrednost mere F1 0,59, na zbirki SALAMI [Smith et al., 2011] pa 0,49.

Podatkovno pretokovni pristop [Weiss and Bello, 2011] za avtomatsko identifikacijo ponavljajočih vzorcev v glasbi analizira matriko značilnic z uporabo verjetnostne latentne analize, neodvisne od zamika (angl. shift-invariant probabilistic latent component analysis - SI-PLCA). Avtorja uporabita omejitve razpršenosti za avtomatsko identifikacijo števila ponavljajočih vzorcev in njihove dolžine. Predlagan pristop je uporabljen za dekompozicijo s taktom poravnane kromatične predstavitve z namenom ekstrakcije ponavljajočih harmoničnih motivov in oceno njihovega položaja v pesmi. Delovanje metode je ovrednoteno na popularni glasbi. Rezultati so primerljivi s pristopom [Weiss and Bello, 2010].

2.3.4 *Verjetnostni modeli za segmentacijo*

Za segmentacijo so raziskovalci uporabljali tudi verjetnostne modele. Avtorja v delu [Paulus and Klapuri, 2009a] predstavita pristop za pridobitev sekcijske predstavitve glasbe s segmentacijo in označbami, kot so refren in kitica. Predstavljena metoda uporablja kromatične, ritmične in barvne lastnosti v zvočnem posnetku. Analiziran posnetek razdeli na veliko možnih kandidatnih odsekov, med katerimi se nato izračunajo podobnostne vrednosti, ki se nato pretvorijo v verjetnostni prostor. Med seboj se združijo vrednosti različnih lastnosti in se uporabijo pri izračunu mere podobnosti za posamezni kandidatni strukturni del. Mera zajema tudi muzikološko predznanje. Na koncu se s pomočjo novega iskalnega pristopa za iskanje poti v usmerjenem acikličnem grafu, kjer posamezno vozlišče predstavlja mogoč segment v signalu, izračuna segmentacija, ki maksimizira mero podobnosti. Pristop je bil ovrednoten na več zbirkah z več kot 550 ročno označenimi popularnimi skladbami. Na zbirki skupine The Beatles dosega vrednost mere F1 0,61, na zbirki RWC-Pop [Goto et al., 2002] 0,65, na zbirki TUTstructure07 [Paulus and Klapuri, 2009b] pa prav tako 0,65.

2.3.5 *Uporaba teorije grafov in gručenja pri segmentaciji*

Med drugim so se nekateri avtorji za segmentacijo glasbe zgledovali tudi po teoriji grafov. Takšen pristop predstavijo avtorji v delu [Panagakis et al., 2011], kjer je graf skonstruiran iz razpršene predstavitve vektorjev značilnic. Avtorji segmentacijo dosežejo z uporabo spektralnega gručenja (angl. spectral clustering) nad dobljenim grafom. Rezultati so obetavni in primerljivi z ostalimi segmentacijskimi tehnikami. Metoda na zbirki popularne glasbe skupine The Beatles dosega vrednost mere F1 0,61.

Teorijo grafov pa uporabita tudi avtorja v [McFee and Ellis, 2014a], kjer se avtorja za segmentacijo ne opirata na samopodobnostne matrike, ampak uporabita spektralno teorijo grafov (angl. spectral graph theory). Le-ta producira nizko-dimenzionalno enkodiranje ponavljajočih struktur in tako izrazijo hierarhične povezave med posameznimi strukturnimi komponentami. Ista avtorja v delu [McFee and Ellis, 2014b] predstavita drugačen pristop, kjer se opirajo na zaporedno linearno diskriminantno analizo (angl. ordinal linear discriminant analysis) z namenom, da se naučijo bodočih projekcij za izboljšavo gručenja časovne vrste. Prav tako predstavita latentne značil-

nice strukturnih ponovitev (angl. latent structural repetition features), ki zagotavljajo fiksno-dimenzionalno predstavitev globalne strukture pesmi in omogočajo tudi modeliranje preko več pesmi. Metodi na zbirki popularne glasbe skupine The Beatles dosegata vrednost mere F1 0,69 in 0,66, na zbirki SALAMI pa 0,55 in 0,51.

Uporaba dvodimenzionalne Fourierove transformacije za gručenje z namenom iskanja meja med posameznimi segmenti v zvočnih posnetkih predstavita avtorja v [Nieto and Bello, 2014]. Z uporabo magnitude izračunanih koeficientov kromatičnih lastnosti se poenostavi problem gručenja, saj so le-ti neodvisni glede na fazo in tonaliteto. Avtorja raziščeta več strategij za določitev meja med segmenti in uporabita metodo *k*-središč (angl. *k*-means) za določitev njihove oznake. Metoda na glasbeni zbirki skupine The Beatles dosega vrednost mere F1 0,77.

2.3.6 Ostali pristopi k segmentaciji

Učinkovit pristop za odkrivanje ponavljajočih vzorcev in strukturne analize glasbe predstavijo avtorji v [Lu et al., 2004], kjer za izračun značilnic uporabijo konstantno Q transformacijo (angl. constant Q-transform - CQT) in na tem definirajo mero podobnosti. Iz značilnic je izračunana samopodobnostna struktura, iz katere s predstavljeno metodo pridobijo ponavljajoče vzorce. Na podlagi pridobljenih ponovitev je nato izvedena analiza glasbene strukture z uporabo nekaj hevrističnih pravil.

V delu [Paulus and Klapuri, 2006] avtorja predstavita pristop, ki na podlagi različnih lastnosti zvočnega posnetka (ritem, takt, kromatične lastnosti in barvne lastnosti) in na podlagi predstavljene cenitvene funkcije z združevanjem posameznih neprekrivajočih delov posnetka izračuna optimalno strukturo zvočnega posnetka, pri čemer za primerjanje izbranih delov uporablja dinamično ukrivljanje časa (angl. dynamical time warpping - DTW), izbiro značilnic pa pristop opravi na podlagi ocene optimalnosti. Pristop na zbirki popularne in rock glasbe dosega vrednost mere F1 do 0,78.

Segmentino [Mauch et al., 2009] je segmentacijska metoda, ki je bila razvita za izboljšanje rezultatov transkripcijskega pristopa. Temelji na iskanju poti v samopodobnostni matriki, izračunani za s taktom sinhronizirane kromatične značilnice. Rezultati algoritma izredno varirajo glede na to kakšen tip glasbe podamo kot vhod. Metoda se na instrumentalu obnese odlično, kar pa ne velja za petje.

Splošen pristop za iskanje meja v časovnih vrstah [Serrà et al., 2012] je bil apliciran tudi na področje segmentacije glasbe z namenom ugotavljanja in označevanja hierarhične strukture glasbe [Serrà et al., 2014]. Metoda imitira človeški kratkočasovni spomin z enkapsulacijo nedavnega dela posnetka, s čimer zaobjame homogenosti in ponovitve tako, da medsebojno primerja pare delov posnetka. Nadalje izračuna strukturne značilnice in razlike teh značilnic, kar predstavlja mero novitete. Vrhovi v grafu mere novitete nakazujejo meje med posameznimi deli v posnetku. Metoda na zbirki popularne glasbe skupine The Beatles doseže vrednost mere F1 0,77, na zbirki RWC-pop pa 0,8. Peeters in Bisot sta predstavljeno metodo združila z metodo [Goto, 2006] za uspešno detekcijo refrenov v zvočnih glasbenih posnetkih [Peeters and Bisot, 2014]. Metoda na istih zbirkah doseže vrednost mere F1 0,78 in 0,76.

Pristop [Ullrich et al., 2014] predstavi uporabo konvolucijskih nevronskih mrež (angl. convolutional neural networks - CNN) na spektalnih značilnicah in samopodobnostnih matrikah, izračunanih iz zvočnih značilnic, za zajetje informacije o glasbeni strukturi. Prav tako avtorji z namenom zajema hierarhične strukture nevronske mreže učijo na dvostopenjskih označbah. Posledično tudi predstavljen algoritem vrne dvostopenjske označbe. Predstavljen pristop ovrednotijo na zbirki popularne glasbe SALAMI. Na tekmovanju MIREX 2013 metoda doseže vrednost mere F1 0,62.

–

Vsi predstavljeni segmentacijski pristopi so bili razviti za segmentacijo popularne ali klasične glasbe. S tem predpostavijo, da je bila glasba profesionalno posneta in vsebuje minimalno stopnjo šuma. Prav tako je izvedena s strani profesionalnih glasbenikov, ki so sposobni točne izvedbe. Veliko pristopov se dodatno opira na omejitve, kot je prisotnost močnega ritma (npr. pri izračunu z ritmom poravnanih zvočnih značilnic) ali skoraj konstantnega tempa.

V prvem delu disertacije prikažemo, da pristopi, zasnovani za segmentacijo popularne ali klasične glasbe, odpovedo na zbirki ljudske glasbe, saj za ljudsko glasbo ne veljajo enake predpostavke. Naš namen je nasloviti specifični problem segmentacije zvočnih posnetkov ljudske glasbe. Število digitaliziranih zbirk ljudske glasbe se zaradi vlaganja v ohranjanje kulturne dediščine hitro povečuje, zaradi česar se povečuje tudi potreba po zanesljivih, prilagodljivih in robustnih metodah MIR, razvitih za ljudsko glasbo.

Posnetki ljudske glasbe predstavljajo kar nekaj izzivov: posnetki so tipično šumni, saj so posneti na terenu v vsakdanjih pogojih; večinoma so izvedeni s strani amaterskih izvajalcev, zaradi česar lahko vsebujejo netočno petje, drsenje višine tonov, pozabljeno besedilo, prekinitve, velike razlike v tempu in podobne nepravilnosti. Velike razlike pa so tudi v sestavih izvajalcev v ljudski glasbi. Tako lahko določene pesmi izvajajo solo pevci, nekaj pevcev ali zbor, lahko gre za instrumentalno glasbo ali za kombinacijo petja in instrumentalne glasbe. Po drugi strani za samo strukturo ljudske glasbe velja, da ni kompleksna. Pesmi so večinoma sestavljene iz ponavljajočih delov, ki imajo isto melodijo, kar olajša avtomatsko segmentacijo. V nadaljevanju sledi pregled prispevkov s področja segmentacije ljudske glasbe.

2.3.7 Segmentacija ljudske glasbe

Enega izmed prvih pristopov k segmentaciji ljudske glasbe predstavijo avtorji v [Müller et al., 2009]. Pristop pričakuje poleg zvočnega posnetka tudi simbolično predstavitev enega ponavljajočega dela kot vzorec, na podlagi katerega nato išče podobnosti v zvočnem posnetku. Segmentacija se izvede s pomočjo izračuna mere razdalje, ki temelji na dinamičnem ukrivljanju časa (angl. dynamic time warping - DTW). Pristop poravna kromatične značilnice (angl. F0-enhanced chroma energy normalized statistics - CENS [Müller, 2007]), izračunane iz simbolične predstavitve, z istimi značilnicami, izračunanimi na zvočnem posnetku. Dinamično ukrivljanje časa skrbi za toleranco v spremenljivem tempu, krožna rotacija kromatičnih značilnic pa tolerira drsenje v višini tonov. Pristop je bil razvit za specifičen tip posnetkov, ki vsebujejo solo petje in se posledično na ostalih ne obnese enako dobro. Metoda doseže vrednost mere F1 0,91 na delu zbirke Onder de groene linde (OGL) [Müller et al., 2010].

Verjetnostni pristop za segmentacijo zvočnih posnetkov, a na višjem nivoju, predstavi Marolt v delu [Marolt, 2009], kjer s predstavljenim verjetnostnim modelom, temelječim na skritih markovskih modelih, segmentira dolge terenske posnetke ljudske glasbe v dele, ki predstavljajo zaključeno enoto glede na tip signala (npr. govor, solo petje, večglasno petje, inštrumentalna glasba ipd.), in posamezne dele tudi označi. Verjetnostni model je naučen na množici posnetkov posameznih enot in nato ovrednoten na množici posnetkov iz zbirke Etnomuza.

V razširitvi obstoječe metode [Müller et al., 2011], ki ne potrebuje več predznanja,

avtorji vpeljejo novo mero ujemanja, ki zna upoštevati tudi večja odstopanja v tempu, instrumentaciji in modulaciji znotraj posameznih ponavljajočih delov kot tudi med različnimi ponovitvami delov glasbenega posnetka. Tudi za ta pristop velja, da deluje dobro na solo petju, na ostalih tipih signala pa precej slabše. Metoda je bila ovrednotena na zbirki solo posnetkov ljudske glasbe, kjer dosega vrednost mere F1 do 0,86.

Drugo adaptacijo iste metode, ki prav tako ne potrebuje predznanja, smo predstavili v [Bohak and Marolt, 2012]. Ta pristop zajema detekcijo vokalnih pavz v posnetku, ki jih uporabimo kot kandidate za začetke ponovitev. Kot primer ponavljajočega vzorca vzamemo prvi del posnetka, ki ga nato s pomočjo dinamičnega ukrivljanja časa, in s tem definirano mero ujemanja, primerjamo z deli posnetka, ki se začnejo z vokalnimi pavzami. Za detekcijo vokalnih pavz je uporabljen pristop [van Kranenburg and Tzanetakis, 2010].

Avtorji [Müller et al., 2010] predstavijo avtomatsko analizo izvedbe posnetih glasbenih materialov s področja ljudske glasbe. Z uporabo več modalnosti avtorji izkoristijo obstoj simboličnega zapisa melodije idealizirane kitice za analizo celotnega posnetka v številnih ponovitvah. Avtorji predstavijo uporabo kromatskih predlog za zaznavanje konsistenc in nekonsistenc v posameznih ponovitvah.

Nov pristop za segmentacijo ljudske glasbe [Müller and Grosche, 2012] uporablja izboljšane samopodobnostne matrike in novo funkcijo ujemanja, razvito na podlagi dinamičnega programiranja. Pristop tolerira tako spremembe v zamikih višine tonov kot v tempu in je bil ovrednoten tako na zbirki popularne in klasične glasbe kot tudi na zbirki ljudske glasbe. Rezultati so obetavni, še posebej zaradi tega, ker gre za pristop, ki pokriva širok nabor glasbe. Uporaba takšne metode za iskanje glasbenih izvlečkov je predstavljena v delu [Müller et al., 2013] in dosega vrednost mere F1 0,87 na delu zbirke OGL.

–

Dobra segmentacija glasbenih zvočnih posnetkov velikokrat predstavlja osnovo za druge naloge. Ena takšnih nalog je tudi glasbena transkripcija. Uspešna segmentacija v ponavljajoče vzorce nam tako omogoča, da transkribiramo zgolj eno ponovitev, ali da za boljšo transkripcijo uporabimo informacijo ostalih ponovitev in tako pridobimo boljši končni rezultat. Opis pristopov za transkripcijo in njene aplikacije podajamo v

nadaljevanju.

2.4 Transkripcija glasbe

Transkripcija glasbe je postopek, pri katerem iz zvočnega posnetka izdelamo simbolični zapis glasbe z namenom kasnejše reprodukcije in/ali analize oz. interpretacije glasbenega dela. Transkripcijo večinoma izvajajo izkušeni glasbeniki z dobrim posluhom, ki so zmožni oceniti višine tonov in trajanja zgolj na podlagi poslušanja posnetka. Kot končni rezultat transkripcije tako v večini primerov pričakujemo notni zapis signala.

Lažji problem predstavlja transkripcija monofoničnih posnetkov, kjer je v posnetku prisoten največ en glas. Dosti težja pa je transkripcija večglasja oz. polifonije, kjer v istem zvočnem posnetku nastopa več glasov in/ali instrumentov. Težavnost transkripcije je seveda odvisna tudi od same kvalitete zvočnega posnetka in od kvalitete same izvedbe.

Glasbena transkripcija se dostikrat uporablja z namenom kodiranja glasbe, po drugi strani pa je povezana tudi s področjem *glasbene percepcije* [Deutsch, 1982]. Zaznavanje in prepoznavanje posameznih zvokov v glasbi predstavlja velik del njene percepcije. Seveda pa se moramo zavedati, da je glasbena notacija primarno razvita z namenom glasbene reprodukcije in ne modeliranju tega, kako človek glasbo sliši. Transkripcija poleg zapisu glasbe velikokrat služi tudi drugotnim namenom: pridobivanju informacij iz glasbe na podlagi melodije, glasbenemu procesiranju, kot je prilagajanje instrumentacije, aranžmaja, prilagajanje glasnosti posameznega dela, za namene razvoja z glasbo povezane opreme, kot je sinhronizacija luči z vzorci v glasbi, interaktivni glasbeni sistemi ipd., glasbeni analizi improvizirane in ljudske glasbe, katerih zapisi redko obstajajo in tudi drugje.

2.4.1 Začetki avtomatske transkripcije

Prvi poskusi avtomatske transkripcije glasbe segajo vse v 70-ta leta 20. stoletja z delom [Moorer, 1975, 1977], kjer avtor predstavi sistem za avtomatsko transkripcijo dvoglasnih kompozicij. Delo Moorerja so nadaljevali [Chafe et al., 1985; Piszczalski, 1986] in [Maher, 1989, 1990] v 80-ih letih 20. stoletja. V vseh predstavljenih zgodnjih sistemih

je bilo število glasov omejeno na dva, prav tako pa so sistemi vsebovali precejšnje omejitve glede spreminjanja višine tonov posameznega glasu. Tudi na področju sledenja ritma, ki je pri transkripciji zelo pomemben, je bilo veliko narejenega v istem obdobju, kar je na kratko povzeto v delu [Lee, 1991]. Pri sledenju ritmu je izredno pomembna tudi transkripcija tolkal, katere prvi poskusi so predstavljeni v [Schloss, 1985; Bilmes, 1993]. Transkripcijo polifoničnih posnetkov tolkal predstavita avtorja v delu [Goto and Muraoka, 1994]. Podrobnejša predstavitev začetkov transkripcije je predstavljena v [Tanguiane, 1993].

2.4.2 *Uporaba statističnih metod*

Veliko raziskovalcev je za reševanje problema transkripcije poseglo po statističnih metodah. Avtorji v [Kashino et al., 1995] tako zaznavno organizacijo zvoka predstavijo kot problem analize scene v zvočni domeni. Predstavljen model temelji na večprocesnih modulih in mreži hipotez za kvantitativno integracijo informacij različnih virov. Model je asinhron in ob posameznem dogodku informacije avtomatsko integrira v mrežo hipotez, ta pa zgradi optimalni model za opis zaznanega zvoka.

Goto v delu [Goto, 2001] opiše robustno metodo za detekcijo osnovne frekvence melodije in basovske linije. Metoda nima vnaprej določenega števila zvočnih virov, ampak avtomatsko določi najbolj dominantno melodijo in basovsko linijo. Metoda oceni relativno pomembnost posamezne melodične linije s pomočjo verjetnostne funkcije in oblikuje ton na podlagi ocene maksimalne aposteriorne verjetnosti. Rezultati nakazujejo, da lahko metoda v realnem času robustno zazna melodijo in basovsko linijo. Predstavljen pristop pravilno zazna vodilno melodijo v 88,4 % primerov in basovsko linijo v 79,9 % primerov na zbirki RWC [Goto et al., 2002].

Pristop [Godsill and Davy, 2002] za transkripcijo uporablja zgodnji Bayesov model za opis posameznih komponent signala: osnovne frekvence, delnih tonov in amplitude. Osnovni model je prilagojen za bolj resnične posnetke tako, da pokriva tudi *ne-beli* del spektra, časovno spreminjajočo se amplitudo in netočne delne tone. Neznani parametri modela so simulirani z algoritmom MCMC (angl. Markov chain Monte Carlo). Model je precej splošen, njegova uporaba za namene transkripcije zgolj ena izmed možnih.

Bayesov harmonični model za oceno višine tonov je predstavljen v delu [Davy and Godsill, 2003], kjer avtorja predstavita uporabo Bayesovih hierarhičnih struktur z namenom ocene količin, kot so višina tonov, glasbena dinamika, barva zvoka ipd. Predstavljen model avtorja uporabita za glasbeno transkripcijo, model pa je omejen na transkripcijo nezvenečih instrumentov z omejenim številom sočasnih zvočnih virov.

Avtorji v [Kameoka et al., 2004] za ločevanje harmoničnih struktur uporabijo model mešanih Gaussovih verjetnostnih porazdelitev (angl. Gaussian mixture model - GMM), s katerim modelirajo posamezno harmonično strukturo. Predstavljen pristop omogoča oceno števila in oblike podležnih harmoničnih struktur na podlagi največje ocene verjetnosti parametrov modela z algoritmom maksimizacije pričakovanja (angl. expectation-maximization algorithm - EM) in informacijskih kriterijev. Pristop deluje neodvisno od števila in vrste zvočnih virov in vrne točne osnovne frekvence z uporabo enostavnih pristopov v spektralnem prostoru.

Metoda, predstavljena v [Ryynänen and Klapuri, 2005], je bila prav tako razvita z namenom delovanja na realnih zvočnih posnetkih. Je neobčutljiva na prisotnost tolkal, a jih ne transkribira, kot izhod pa vrača predstavitev MIDI. Pristop uporablja skrite markovske modele za opis zvočnih dogodkov, ki na podlagi treh značilnic, izračunanih s pomočjo ocenjevalnika osnovnih frekvenc, izračuna verjetnosti posameznih not in izvede njihovo časovno segmentacijo. Prehodi med posameznimi notami so izvedeni s pomočjo muzikološkega modela. Avtorji so izvedli evalvacijo na množici realnih posnetkov in prišli do vzpodbudnih rezultatov. Metoda dosega vrednost priklica 0,39 in vrednost natančnosti 0,41.

2.4.3 Uporaba računskih modelov

Kako deluje človeški slušni sistem, opisujejo tudi različni računski modeli človeškega slušnega sistema (angl. computational models of the human auditory system). Takšni sistemi so bili razviti z različnimi nameni. Razviti so bili tudi modeli za namene transkripcije.

Delo [Martin, 1996] predstavlja dvodelni sistem za transkripcijo polifonične glasbe. Ozadni del sistema je razvit na podlagi integracije znanja ekspertov in uporabe procesiranja signalov v skladu s predvidenim delovanjem človeškega slušnega sistema. Avtorji

takšen odzadnji model združijo z osprednim modelom, ki temelji na percepciji višine tonov na podlagi logaritmično zakasnjenega korelograma (angl. log-lag correlogram). Rezultati nakazujejo na zmanjšanje oktavnih napak v preprosti polifonični glasbi.

Model za računsko učinkovito analizo polifonije in periodičnosti kompleksnih zvočnih signalov [Tolonen and Karjalainen, 2000] razdeli signal na dva dela glede na frekvenco, izračuna posplošeno avtokorelacijo spodnjega dela signala in ovojnice zgornjega dela ter ju sešteje v enotno avtokorelacijsko funkcijo, ki je v nadaljnjem koraku še dodatno izboljšana. Na podlagi pridobljenih avtokorelacijskih funkcij (osnovne in izboljšane) metoda nato analizira periodične ponovitve signala. Metoda je izredno hitra in primerljiva s pristopi, ki delujejo v časovni domeni.

Metoda za oceno osnovnih frekvenc več sočasnih glasbenih zvokov [Klapuri, 2005] je sestavljena iz računskega modela človeškega slušnega sistema in iz mehanizma za analizo periodičnosti. Detekcija več osnovnih frekvenc je izvedena tako, da je zaznan ton iz signala odstranjen, nato je postopek ponovljen v preostalem signalu. Sama zahtevnost metode ni velika, saj je osnovni računski model izračunan samo enkrat.

Med računske modele prištevamo tudi pristope, ki modelirajo človeško analizo zvočne scene (angl. human auditory scene analysis) [Bregman, 1990]. Tako je v delu [Mellinger, 1991] predstavljen pregled psiho-akustičnih nevro-psiholoških študij, ki raziskujejo, kako je lahko človeški slušni sistem sposoben zaznati izvore posameznih glasov v okolju, kjer je več zvočnih virov. Delo prav tako oriše arhitekturo sistema, ki modelira zgodnji del zaznavanja zvoka v človeškem slušnem sistemu. Ta sistem združuje posamezne lokalne značilnice v posamezne zvočne dogodke in skupine dogodkov ter nazadnje v posamezne zvočne vire. Tako razvit model vsebuje tudi filtre za posamezne zvočne frekvence in je tako sposoben tudi transkripcije.

Avtorja v [Kashino and Tanaka, 1993] opisujeta sistem za ločevanje izvorov zvoka s sposobnostjo modeliranja posameznih tonov. Predstavljen sistem za vhod vzame enokanalni zvočni signal, ki vsebuje zvoke različnih instrumentov. Vhod loči glede na posamezne vhodne instrumente in zgradi MIDI predstavitev, kjer je na posameznem MIDI kanalu transkribiran posamični instrument.

Model, predstavljen v [Godsmark and Brown, 1999], je zasnovan z namenom boljše integracije kazalcev različnih virov v končni izhod modela. Avtorji pokažejo, kako

lahko predstaljen model dobro ponazori poslušalčevo percepcijo prepletenih melodij in hkrati loči različne linije v polifoničnih posnetkih.

Sistem [Sterian and Wakefield, 1999] za sledenje glasbenih delov v polifoničnih posnetkih uporablja hierarhijo časovno-frekvenčnih Kalmanovih filtrov za ugotavljanje pričetkov in koncev posameznih melodičnih delov v zvočnem posnetku.

Pristop, ki je bil primarno razvit za transkripcijo klavirske glasbe [Marolt, 2004], za transkripcijo uporablja model mreže adaptivnih oscilatorjev v povezavi z akustičnim modelom. Pristop na zbirki klavirske glasbe odkrije preko 90 % pravilnih not, okoli 30 % pa je not, ki jih ni v originalnem signalu.

2.4.4 Uporaba metod za ločevanje zvočnih izvorov

Raziskovalci so za namene transkripcije uporabili tudi nekatere izmed metod za analizo neodvisnih komponent (angl. independent component analysis - ICA). Metoda, predstavljena v [Lepain, 1999], uporablja več modelov za identifikacijo posamezne zaznane višine tona, izražene z različnimi spektralnimi porazdelitvami. Avtor metodo preizkusi na širokem naboru zvočnih signalov in prikaže različne možnosti za uporabo same metode.

Pristop, predstavljen v [Smaragdis, 2001], predstavi matematične principe za združevanje nižjeležečih *slušnih* funkcij z namenom formuliranja globalne teorije računskega poslušanja. Model vključuje še nekatere nizkonivojske koncepte, ki simulirajo delovanje notranjega ušesa v obliki filtrov in hevristik. Avtor je pristop preizkusil za namen ločevanja zvočnih virov na realnih zvočnih posnetkih.

Delo [Abdallah, 2002] predstavi raziskavo principov za zmanjšanje redundance s pomočjo nenadzorovanega učenja za predstavitev zvoka in glasbe. Avtor predstavi ogrodje za učenje z verjetnostnim modelom, ki ga aplicira na dve nizkonivojski predstavitvi, kar nadalje aplicira za namen transkripcije polifonične glasbe. Izboljšan pristop predstavita avtorja v [Abdallah and Plumbley, 2004]. Podoben pristop za polifonično transkripcijo tolkal in ugotavljanje zvočnih virov predstavi avtor v [Fitzgerald, 2004], kjer združi pristop ISA z dodatnim predznanjem. Nadalje prikaže tudi uporabo metode analize posamičnih podprostorov (angl. prior subspace analysis - PSA) za polifonično transkripcijo tolkal. Transkripcijo tolkal z uporabo metode NMF predstavita avtor-

ja v delu [Paulus and Virtanen, 2005]. NMF je v pristopu uporabljena za izločitev posameznega zvočnega vira tolkal iz zvočnega posnetka, kar je nadalje uporabljeno za detekcijo začetkov posameznih zvočnih dogodkov. Sistem je bil preizkušen in daje do 96 % točne zadetke v transkripciji tolkal.

2.4.5 Sodobnejši pristopi k transkripciji

Poglobljen pregled področja transkripcije do leta 2006 je predstavljen v delu [Klapuri, 2006], kjer avtor poleg pregleda dotedanjih pristopov predlaga tudi lastne pristope za reševanje problemov povezanih s transkripcijo glasbe [Klapuri and Davy, 2006]. Pregled novejših pristopov k transkripciji do leta 2012 predstavijo avtorji v [Grosche et al., 2012]. Večina sodobnih pristopov temelji na spektralni dekompoziciji (NMF, PLCA ipd.).

Pristop [Cont, 2006] se od ostalih razlikuje v tem, da namesto izločanja profilnic višin tonov iz signala spekter signala projicira na vnaprej določene predloge višin tonov. Dekompozicijski algoritem temelji na metodi NMF. Metoda na majhni glasbeni zbirki presega natančnost 0,7. Nekoliko novejši pristop s podobno zasnovo predstavijo avtorji v [Dessein et al., 2010]. Prednost predstavljenega sistema je v tem, da je sposoben transkribirati glasbo v realnem času. Na uporabi sorodne metode - metode nenegativne matrične aproksimacije (angl. non-negative matrix approximation - NNMA) temelji tudi pristop, predstavljen v [Raczyński et al., 2007]. Za razliko od sorodnih pristopov ta v temeljni matriki vsebuje zgolj harmonične lastnosti. To v kombinaciji z dodatnim kaznovanjem skupnih nepodobnosti in razpršenosti vrstic poskrbi, da metoda vrača boljše rezultate od dotedanjih pristopov. Na predstavljeni glasbeni zbirki presega natančnost 0,7. [Niedermayer, 2008] predstavi pristop, ki metodo NMF uporablja za transkripcijo polifonične glasbe enega samega instrumenta (npr. klavirja). Avtor predlaga uporabo fiksnega nabora bazičnih vektorjev, ki modelirajo posamezne tone instrumenta. Metoda na solo klavirski glasbi dosega vrednost mere F1 0,52. Avtorji v delu [Boulanger-Lewandowski et al., 2012] prav tako uporabljajo metodo NMF za transkripcijo. V prispevku podajo možne izboljšave z uporabo dodatnih diskriminativnih kriterijev za povečanje stopnje razpršenosti značilnic različnih razredov. Na dveh glasbenih zbirkah dosega vrednost mere F1 preko 0,61. Transkripcijska metoda [Bay et al., 2012] se zgleduje po predhodnih pristopih in je razvita z namenom sledenja vi-

šini tonov posamičnih instrumentov ali glasov v polifoničnih zvočnih posnetkih. Pri tem se zanaša na verjetnostne lastnosti razdelitve spektra na posamezne dele ter na note posameznih instrumentov iz vnaprej sestavljene zbirke. Metoda dosega vrednost mere F1 na zbirki MIREX Bach 0,55. Isti avtor je predstavil tudi sistem za ovrednotenje različnih transkripcijskih sistemov [Bay et al., 2009].

Avtorji v [Cheng et al., 2013] predlagajo razširitev uporabe metode PLCA za transkribcijo z uporabo deterministočnega ohlajanja z EM. S tem želijo avtorji preprečiti, da se metoda PLCA ustavi v nekem lokalnem optimumu in želijo dopustiti možnost za iskanje boljšega optimuma ter s tem preseči rezultate metod, ki uporabljajo NMF ali PLCA. Metoda na zbirki MIREX doseže vrednost mere F1 0,55. Delo [Grindlay and Ellis, 2010] opisuje transripcijski sistem, ki ne potrebuje predznanja o instrumentih, prisotnih v zvočnem signalu, a lahko takšno informacijo dodatno izkoristi. Predlagan sistem je sposoben modeliranja posamičnega instrumenta, prisotnega v signalu, in je tako sposoben posamezne note pripisati določenemu instrumentu. Sistem se uči na vnaprej definirani zbirki zvokov posameznih instrumentov, na podlagi česar se priuči modela, ki ga kasneje uporabi za transkripcijo. Na sintetizirani zbirki Bacha dosega vrednost mere F1 0,53. Metoda [Fuentes et al., 2013] za transkripcijo uporablja izvedenko metode PLCA - metodo harmonično adaptivne latentne analize komponent (angl. harmonic adaptive latent component analysis), s katero modelirajo časovne variacije tako v domeni višine tonov kot v spektralni domeni. Metoda je bila ovrednotena na dveh glasbenih zbirkah z vrednostjo mere F1 0,31.

Sistem za transkripcijo specifičnih zvočnih posnetkov - igranja ur z zvonovi - je predstavljen v delu [Marolt and Lefeber, 2010]. Gre za povsem specifične instrumente, zaradi česar za njihovo transkripcijo obstoječi pristopi niso primerni. Pristop temelji na verjetnostnem modelu, ki maksimizira skupno verjetnost sekvenc notnih zaporedij. Pristop za transkripcijo pritrkavanja cerkvenih zvonov [Marolt, 2012] oceni število zvonov v posnetku, nato pa s pomočjo metod k-središč in NMF transkribira posnetek. Pristop za nenadzorovano transkripcijo klavirske glasbe [Berg-Kirkpatrick et al., 2014] sestoji iz treh verjetnostnih modelov: dogodkovnega modela, aktivacijskega modela in spektralnega modela, na podlagi katerih sestavi končno transkripcijo klavirske glasbe.

Benetos je s soavtorji predstavil več pristopov. Prvi je predstavljen v delu [Benetos and Weyde, 2013], kjer avtorja predstavita uporabo markovskih modelov z določeni-

mi časovnimi koraki za modeliranje posamezne višine tona. Ti modeli so vključeni v konvolucijsko verjetnostno ogrodje za modeliranje časovnega razvoja in trajanj posameznih zvokov. Rezultat je dvostopenjski transkripcijski postopek, ki vključuje sledenje notam in zagotavlja robustnejše delovanje. Naslednji pristop [Benetos and Dixon, 2013] temelji na uporabi na zamik neodvisne metode PLCA (angl. shift-invariant probabilistic latent component analysis). Kot v prejšnji metodi, so tudi tukaj v uporabi skriti markovski modeli za sledenje notam. Dodatna razširitev je predstavljena v delu [Benetos et al., 2014]. Razširitev poleg detekcije nezvenečih instrumentov omogoča tudi detekcijo zvenečih instrumentov iz nabora tolkal in njihovo transkripcijo. Dodatno prilagojen sistem [Benetos and Weyde, 2015] predlaga uporabo spremenljive Q-Transformacije (angl. variable Q-transform), namesto pogosto uporabljene Q-transformacije, za predstavitev frekvenčnega spektra signala, s čimer še dodatno izboljša rezultate sistema.

Razvitih in predstavljenih je bilo veliko transkripcijskih algoritmov, ki pa na realnih zbirkah ne delujejo dobro. To je razvidno tudi iz rezultatov evalvacije MIREX 15, kjer na zbirki realnih instrumentalnih posnetkov (zbirka Su [Su and Yang, 2015]) metode dosegajo vrednost mere F1 0,57. Transkripcijskih metod za ljudsko glasbo je malo. Takšen primer je metoda za ekstrakcijo melodije iz posnetkov flamenka [Salamon and Gómez, 2012] ali metoda za transkripcijo turške mikrotonalne glasbe [Benetos and Holzapfel, 2015]. Obstoječe splošne transkripcijske metode na ljudski glasbi odpovedo, kar pokažemo tudi v nadaljevanju.

–

Rezultati transkripcije so uporabni že sami po sebi, saj omogočajo reprodukcijo glasbe s strani drugih izvajalcev. Poleg tega je transkripcija lahko tudi podlaga za nadaljnje korake pri obdelavi glasbe. V našem primeru rezultate transkripcije uporabimo za to, da iz celotne transkripcije izločimo tisti del, ki najbolje odraža vse ostale dele. Ravno iskanje najbolj reprezentativnega dela pesmi oz. skladbe pa je predstavljeno v nadaljevanju.

2.5 Iskanje najbolj reprezentativnega dela glasbe

Iskanje najbolj reprezentativnega dela glasbe si lahko predstavljamo na več načinov. Največkrat to pomeni tisti del glasbenega zvočnega posnetka, ki poslušalca spomni na predvajano pesem. To je lahko dostikrat tudi zelo subjektivno, saj je človeški spomin asociativen, povezave pa se pri vsakem posamezniku povezujejo na drugačne načine. Za popularno glasbo, ki nima nujno enostavne strukture, poslušalca na predvajano pesem velikokrat spomni že sama uvodna melodija. Največkrat pa pesem prepoznamo po melodiji refrena. V ljudski glasbi takšnih problemov večinoma ni, saj so pesmi večinoma sestavljene iz melodično ponavljajočih delov.

Del področja MIR, ki se ukvarja s to problematiko, je področje povzemanja glasbe (angl. music summarization oz. audio thumbnailing). Ideja povzemanja glasbe pa ne pomeni nujno iskanje zgolj enega dela zvočnega posnetka ampak tudi več delov, ki dobro povzamejo celoten posnetek. Med prvimi prispevki s področja povzemanja glasbe sta prispevka [Logan and Chu, 2000; Chu and Logan, 2000], ki opisujeta povzemanje glasbe z uporabo ključnih fraz. Za primere rock glasbe, sestavljene iz delov kitica (angl. verse) in refren (angl. chorus), je cilj metode vrniti del z refrenom ali največkrat ponovljenim in posledično najbolj zapomnljivim delom glasbenega posnetka. Pristop je sestavljen iz treh delov: (1) parametrizacija pesmi v značilnice, (2) iskanje strukture na podlagi značilnic s pomočjo gručenja segmentov fiksne dolžine ali z uporabo modela HMM, (3) na podlagi hevristike se izberejo ključne fraze. Na zbirki pesmi skupine The Beatles pristop ovrednotijo in ugotovijo, da je uporaba gručenja boljša od uporabe HMM-ja in izbora naključnih delov.

Avtorja [Cooper and Foote, 2002] predstavita metodo za iskanje glasbenega povzetka, kar za avtorja hkrati pomeni iskanje najbolj reprezentativnega dela. V pristopu za to iščeta del posnetka, za katerega velja, da je stopnja podobnosti s celotnim posnetkom maksimalna. Po parametrizaciji v posamezna okna se med pari oken izračuna kvantitativna mera podobnosti in izračuna dvodimenzionalna matrika podobnosti. Z izračunom vsote mer podobnosti za posamezni segment znotraj podobnostne matrike je izražena mera podobnosti med posamičnim delom in celotnim posnetkom. Avtorja obravnavata tudi nekaj variacij predstavljene metode in predstavita rezultate na glasbenih posnetkih različnih zvrsti.

Iskanje glasbenih povzetkov na podlagi strukturne analize predstavita avtorja v delu [Chai and Vercoe, 2003]. Avtorja predstavita tri strategije za iskanje povzetkov, ki temeljijo na rezultatih strukturne analize. Predlagata tudi izboljšave osnovne metode strukturne analize. Namesto subjektivne evalvacije avtorja rezultate ovrednotita na podlagi lastnosti povzetkov s komercialnih spletnih strani. Na podlagi strukturne analize pa deluje tudi pristop, predstavljen v [Bartsch and Wakefield, 2005]. Cilj metode je na podlagi strukturne redundance v kromatični predstavitvi poiskati del posnetka, ki predstavlja refren.

Metoda [Xu et al., 2005] temelji na gručenju izračunanih značilnic in domenskega znanja glede na *čisto* in vokalno glasbo. Razlikovanje med obema tipoma glasbe je z uporabo izračunanih značilnic izvedeno s pomočjo metode podpornih vektorjev (angl. support vector machines - SVM), ki se izkaže za bolj zanesljivo od uporabe modela HMM ali uporabe evklidske razdalje med značilnicami. Avtorji delovanje metode ovrednotijo tudi na zbirki posnetkov.

Avtomatsko generiranje glasbenih povzetkov je predstavljeno tudi v pristopu [Zhang and Samadani, 2007]. Kot že nekaj predstavljenih metod tudi ta temelji na strukturni analizi. V prvem koraku metoda zazna dele pesmi s ponavljajočo melodijo, v naslednjem koraku identificira dele posnetka, ki vsebujejo vokale, na podlagi teh informacij pa iz posnetka pridobi strukturo. Nazadnje na podlagi hevrističnih pravil izbere enega ali več glasbenih povzetkov glede na glasbeno strukturo. Pristop se izkaže za učinkovitega in je bil preizkušen na pesmih različnih žanrov.

Pristop, predstavljen v delu [Ferguson and Cabrera, 2009], za izračun glasbenega povzetka uporabi spektralno analizo zvočnega signala. Cilj metode je, da posamezni del zvočnega signala izvzame iz časovnega konteksta in ga predstavi kot navidezno stacionarni signal s spektrom, enakim originalnemu. V delu avtorja predstavita možnost uporabe takšne predstavitve signala za izračun glasbenega povzetka.

Pristop [Schuller et al., 2010] predstavi, kako se lahko za izračun glasbenega povzetka uporabi harmonična in ritmična struktura glasbe. Ritmična struktura temelji na ritmu in je izračunana z množico filtrov z neskončnim enotinim odzivom. Za izračun harmonične strukture so uporabljene značilnice, ki modelirajo normalizirane statistike porazdelitve kromatične energije, na podlagi katerih se izračuna podobnostna matrika. Za pridobitev končnega glasbenega povzetka avtorji predlagajo uporabo metod s

področja procesiranja slik.

–

Iskanje glasbenega povzetka ali najbolj reprezentativnega dela glasbenega zvočnega posnetka se izkaže za precej težaven problem. Te problematike na domeni ljudske glasbe do sedaj raziskovalci še niso naslavljali. Vsekakor bi bili rezultati takšnih pristopov na ljudski glasbi interesantni predvsem za hitrejše pregledovanje zbirk ljudske glasbe, nenazadnje pa tudi pri samem iskanju po takšnih zbirkah. Ta ideja je tudi povod za razvoj lastne metode na podlagi rezultatov segmentacije in transkripcije, kar je predstavljeno v drugem delu disertacije.

–

S tem zaključujemo pregled področij, povezanih s tematiko disertacije. V nadaljevanju sledi predstavitev dela, izvedenega v okviru disertacije, razdeljenega v dva glavna vsebinska dela.

Del I

Segmentacija zvočnih posnetkov ljudske glasbe

V prvem delu disertacije bomo predstavili novorazvito segmentacijsko metodo za segmentacijo ljudske glasbe. Za razvoj nove metode smo se odločili zaradi nezadovoljivega delovanja obstoječih metod na zvočnih posnetkih ljudske glasbe, kar pokažemo v naslednjem poglavju. Potreba po razvoju algoritma se je porodila med sodelovanjem z Glasbenonarodopisnim inštitutom Znanstvenoraziskovalnega centra Slovenske akademije znanosti in umetnosti (GNI ZRC SAZU), kjer je anotacija in analiza posnetkov ljudske pesmi del raziskovalnega procesa.

Ker so zbirke ljudskega gradiva velike in se s časom še širijo, je potreba po zagotavljanju različnih možnosti iskanja po zbirkah in njihove analize vse večja. Z dobro segmentacijo dobimo o posameznem posnetku dodatne informacije, kar nam olajša nadaljnjo analizo, tako ročno kot avtomatsko, npr. transkripcijo. Ker rezultati obstoječih segmentacijskih metod niso zadovoljivi, smo se v okviru disertacije odločili razviti lastno segmentacijsko metodo.

Razvita metoda temelji na verjetnostnem modelu, ki maksimizira verjetnost segmentacije zvočnega posnetka. Predstavljeno metodo smo ovrednotili na zbirki ljudske glasbe in njene rezultate primerjali z rezultati aktualnih segmentacijskih metod.

V nadaljevanju najprej predstavimo zbirko zvočnih posnetkov, ki smo jo uporabili za izvedbo evalvacije, nadalje predstavimo ovrednotenje trenutno aktualnih pristopov na predstavljeni zbirki posnetkov, kjer izpostavimo pozitivne in negativne lastnosti posameznega pristopa. Sledi predstavitev osnovnih pojmov, potrebnih za razumevanje razvite metode, predstavitev razvite metode, njeno ovrednotenje na predstavljeni zbirki podatkov in primerjava s trenutno aktualnimi metodami. Za konec prvega dela podamo še možne izboljšave in prilagoditve razvite metode.

Zbirka zvočnih posnetkov - Etnomuza

Poglavitni rezultat projekta *EtnoMuza: digitalna multimedijska shramba slovenske ljudske glasbene in plesne kulture*, med letoma 2006 in 2008, je digitalni arhiv slovenske ljudske glasbe - *Etnomuza*, predstavljen v [Strle and Marolt, 2007]. Pri izdelavi glasbenega arhiva so sodelovali Laboratorij za računalniško grafiko in multimedije Fakultete za računalništvo in informatiko Univerze v Ljubljani in GNI ZRC SAZU. Od vzpostavitve digitalnega arhiva se količina digitaliziranih vsebin v njem povečuje tako zaradi

digitalizacije obstoječega arhiva kot zaradi dodajanja novih vsebin.

V digitalnem arhivu se nahajajo rokopisi, zvočni posnetki, plesi in slikovno gradivo. Med določenimi vsebinami arhiva obstajajo povezave, spet druge pa med seboj niso povezane. Velja na primer, da je večina zvočnih posnetkov povezanih z ustreznimi terenskimi zapisniki, ki vsebujejo metapodatke o zvočnih zapisih. Arhiv vsebuje preko 13.000 rokopisov, preko 1.000 plesov, slikovno gradivo in preko 1.000 terenskih zapisnikov. Za potrebe dodajanja, iskanja po zbirki in pregleda vnosov (rokopisov, zvočnih posnetkov, slik) je bil razvit uporabniški vmesnik. Prav pri iskanju po teh in podobnih zbirkah je odprtih veliko možnosti za integracijo metod področja MIR.

Sestava uporabljenih zvočnih materialov

Ključni materiali v zbirki Etnomuza so terenski posnetki in terenski zapisniki. Le-te sestavljajo etnomuzikologi med svojim terenskim delom, kjer skozi intervjuje s posamezniki odkrivajo značilnosti posnetih pesmi in tradicij. Terenski zapisnik je dokument z informacijami o terenskih posnetkih. Vsebuje podatke o tem, kdaj in kje je posnetek nastal, kdo je njegov avtor, kaj vse je v posnetku, kdo vse so osebe, katerih glasovi so na posnetku, kakšno napravo je zapisovalec uporabljal pri snemanju ipd.

Posamezni terenski posnetek lahko vsebuje različno vsebino; vsebuje lahko pogovore s posamezniki ali skupinami, ljudske pesmi v različnih izvedbah, ki jih izvajajo posamezniki ali skupine, instrumentalno ljudsko glasbo in druge stvari. Pogovori so namenjeni ugotavljanju izvora pesmi, priložnosti, ob katerih se pesmi izvajajo, dokumentiranju podatkov o informatorjih ipd.

Za potrebe disertacije smo iz terenskih posnetkov zbirke EtnoMuza zbrali 159 pesmi različnih sestavov. Poleg teh smo iz zbirke pesmi *Onder de groene linde* (OGL) [Müller et al., 2010] pridobili še 47 pesmi, kar skupno predstavlja 206 pesmi. Zbirka OGL je zbirka nizozemskih ljudskih pesmi instituta Meertens, ki vsebuje večinoma posnetke solo petja. Posamezne pesmi so izvedene s strani različnih sestavov. Tako imamo v zbirki solo petje, dvo- in triglasno petje, zborovsko petje, instrumentalne skladbe in kombinacijo instrumentala in petja. Podrobnejša sestava zbirke je podana v tabeli 2.1.

Kot smo omenili že v uvodu tega dela, je struktura ljudskih pesmi precej preprosta, kar velja tudi za zbrane pesmi. V večini primerov gre za ponavljajoče kitice. Za pe-

Tabela 2.1: Zbirka pesmi, uporabljenih pri razvoju in ovrednotenju predstavljenega pristopa.

Tip	*število pesmi*	*Trajanje (min)*
solo petje (OGL)	47	156
solo petje (Etnomuza)	31	72
dvo- in troglasje (Etnomuza)	30	80
zborovsko petje (Etnomuza)	35	92
instrumental (Etnomuza)	33	74
instrumental in petje (Etnomuza)	30	60
Skupaj	*206*	*534*

smi v predstavljeni zbirki smo dodali ročne anotacije, ki označujejo začetke in konce posameznih ponovitev znotraj posamezne pesmi. Vseh kitic (ponavljajočih delov) v zbirki je 1491, kar v povprečju pomeni 7,24 ponovitev na posamezno pesem. Povprečna pesem je dolga 155 sekund, povprečna kitica pa 26,7 sekunde. Kot zanimivost naj izpostavimo, da ima pesem z največ kiticami kar 34 kitic, najmanj pa dve kitici. Najdaljša pesem je dolga 523 sekund, najkrajša pa 32.

3

Ovrednotenje aktualnih metod segmentacije

Science, however, is never conducted as a popularity contest, but instead advances through testable, reproducible, and falsifiable theories.

– Michio Kaku

Z namenom analiziranja uspešnosti aktualnih segmentacijskih metod na ljudski glasbi smo zbrali nekaj javno dostopnih implementacij segmentacijskih algoritmov. Večina jih je vključena v ogrodje za analizo glasbene strukture (angl. music structure analysis framework - MSAF)[1]. Ovrednotili smo uspešnost delovanja naslednjih algoritmov: MSAF-Foote [Foote, 2000], MSAF-SCluster [McFee and Ellis, 2014a], MSAF-SF [Serrà et al., 2012], MSAF-CNMF3 [Nieto and Jehan, 2013] in MSAF-PLCA [Weiss and Bello, 2011]. Vsak algoritem smo testirali na treh tipih značilnic:

- *MFCC* - Mel-frekvenčni kepstralni koeficienti (angl. Mel-frequency cepstral coefficients) [Mermelstein, 1976], ki zajemajo lastnosti barve zvoka,
- *HPCP* - harmonične profilnice tonskih razredov (angl. harmonic pitch class profiles), ki zajemajo harmonične lastnosti zvoka [Gómez, 2006] in
- *Tonetz* - tonične mreže (angl. tonal centroid features), ki modelirajo tonski prostor in jih v svojem delu omenja že Euler [Euler, 1739].

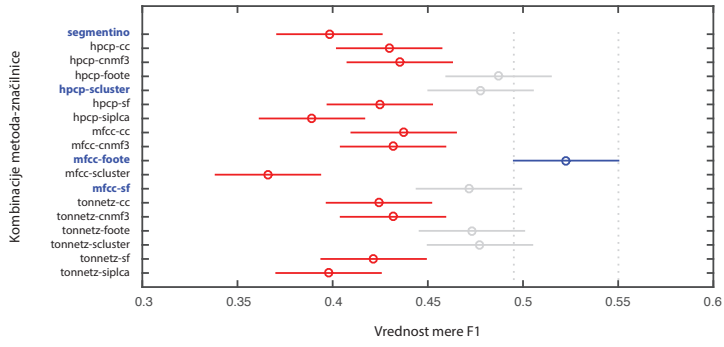
Poleg pristopov, ki so na voljo v omenjenem ogrodju, smo ovrednotili tudi uspešnosti algoritma Segmentino [Mauch et al., 2009].

Vsako izmed teh metod smo ovrednotili na predstavljeni zbirki pesmi in izračunali mero F1. Pri tem ocenjena meja med ponavljajočimi deli velja za pravilno, v kolikor se nahaja v oddaljenosti $\pm$ 3 sekunde od anotacije v posnetku (enako okno je v uporabi na evalvaciji MIREX). Medsebojna primerjava uspešnosti metod glede na mero F1 skupaj s primerjalnimi intervali je prikazana na sliki 3.1 in v tabeli 3.1.

Tabela poleg rezultatov omenjenih algoritmov vsebuje tudi rezultate metode, predstavljene v delu [Müller et al., 2013], ki je bila ovrednotena na delu zbirke Solo (OGL). Implementacija metode ni dosegljiva in je posledično nismo mogli ovrednotiti tudi na preostalem delu zbirke. Podrobnejša obrazložitev rezultatov metode je podana kasneje.

V naslednjih podpoglavjih predstavljamo analizo nekaj najuspešnejših metod.

[1] Ogrodje je dostopno na spletnem naslovu: https://github.com/urinieto/msaf, dostopano 9. junij 2016.



Slika 3.1: Primerjava segmentacijskih metod glede na mero F1 z označenimi signifikantnimi razlikami.

3.1 Segmentino

Segmentacijski algoritem Segmentino [Mauch et al., 2009] je bil originalno razvit za uporabo pri izboljšanju transkripcijske metode. Bazira na iskanju poti v samopodobnostni matriki, izračunani na z ritmom sinhroniziranih kromatičnih značilnicah. Njegova uspešnost je zelo odvisna od vrste sestava v posnetku. Na instrumentalni glasbi se odreže bolje od vseh ostalih metod s povprečno vrednostjo mere F1 0,62, z večjo vrednostjo priklica (angl. recall) in nižjo vrednostjo natančnosti (angl. precision), za kar je razlog nadsegmentacija (angl. oversegmentation). Do nadsegmentacije pride, kadar algoritem meje zazna bolj na gosto, kot so definirane v anotiranih podatkih. Po drugi strani pa je algoritem najmanj uspešen med vsemi na posnetkih solo petja z mero F1 0,36. Razlog za to pa je podsegmentacija (angl. undersegmentation), do katere pride, ko algoritem meje med ponavljajočimi deli postavi redkeje, kot so definirane v anotiranih podatkih.

Glavni razlog za takšno obnašanje algoritma je dejstvo, da algoritem uporablja z ritmom sinhronizirane značilnice pri izračunu samopodobnostne matrike. Takšne značilnice se uporabljajo z namenom obvladovanja sprememb tempa v pesmi, kar pa bazira na predpostavki, da lahko ritem v pesmi zanesljivo ocenimo. Medtem ko je temu lažje zadostiti v primeru instrumentalne glasbe in je algoritem v tem primeru uspešen, pa je prepoznava ritma v posnetkih petja izredno težavna, saj v petju ni ostrih začetkov

Tabela 3.1: Rezultati ovrednotenja izbranih aktualnih metod na zbirki ljudskih pesmi. Mere natančnost (P), priklic (R) in F1 so izračunane kot povprečna vrednost mer za posamezno pesem.

Sestav		Segmentino	MSAF-MFCC-Foote	MSAF-HPCP-SCluster	MSAF-MFCC-SF	Müller
	P	0,86	0,38	0,41	0,36	
Solo (OGL)	R	0,17	0,76	0,51	0,42	
	F1	0,25	0,51	0,46	0,39	0,87
	P	0,82	0,48	0,44	0,52	
Dvo- troglasje	R	0,27	0,85	0,50	0,60	
	F1	0,33	0,61	0,47	0,55	
	P	0,55	0,31	0,38	0,33	
Instrumental	R	0,82	0,97	0,87	0,80	
	F1	0,62	0,47	0,53	0,47	
	P	0,55	0,37	0,35	0,42	
Instr. - petje	R	0,72	0,86	0,64	0,76	
	F1	0,59	0,52	0,46	0,54	
	P	0,92	0,46	0,45	0,46	
Solo	R	0,26	0,78	0,64	0,49	
	F1	0,36	0,57	0,53	0,48	
	P	0,74	0,31	0,40	0,40	
Zbor	R	0,36	0,76	0,61	0,64	
	F1	0,41	0,44	0,48	0,50	
	P	0,74	0,39	0,41	0,41	
Skupaj	R	0,40	0,81	0,59	0,56	
	F1	0,40	0,52	0,48	0,47	

višin tonov, pavze med posameznimi ponovitvami so lahko relativno dolge, prav tako pa lahko v posnetku prihaja do večjih variacij v samem tempu. V kolikor pa odpove detekcija ritma, bo prav tako slab tudi končni rezultat segmentacijskega algoritma. Poleg tega pristop predpostavlja še dodatne omejitve pri iskanju ponovitev (npr. začetki ponovitev se lahko nahajajo na večkratniku štirih udarcev, ponovitve so lahko zgolj določenih dolžin), ki pa za ljudsko glasbo ne držijo v enaki meri kot za popularno glasbo. Prav tako pristop ne naslavlja problema drsenja višine tonov, ki je za ljudsko petje precej običajno.

3.2 MSAF-MFCC-Foote

Foote [Foote, 2000] je predstavil eno prvih metod za segmentacijo glasbe, ki se med testiranimi metodami izkaže za najuspešnejšo na naši glasbeni zbirki. Metoda temelji na meri novitete, izračunani iz samopodobnostne matrike, ki temelji na značilnicah MFCC. Njena uspešnost za različne glasbene sestave je relativno enaka. Vrednosti mere F1 se gibljejo med 0,44 za zborovsko petje in 0,61 za dvo- in triglasno petje. Večinoma je nagnjena k nadsegmentaciji za vse tipe posnetkov.

Glavni razlog za nepravilno postavljene meje predstavlja dejstvo, da značilnice MFCC ne zajemajo harmoničnih glasbenih lastnosti, ampak lastnosti barve zvoka. Ker se barva zvoka v ljudskih pesmih tipično ne spreminja dosti, se to izraža v visoki stopnji samopodobnosti skozi celotno pesem, kar otežuje odkrivanje meja med ponavljajočimi deli. Poudarek je prav tako na razlikovanju med posameznimi vokali (na primer med "A" in "O"), kar prav tako ni relevantno, saj se besedilo v ljudskih pesmih večinoma ne ponavlja. Za povrh pa je metoda zelo občutljiva na velikost jedra, ki se uporabi pri izračunu mere novitete in je to velikost težko uganiti.

Uporaba značilnic MFCC je hkrati tudi razlog, zakaj se metoda obnese relativno dobro. Premori v posnetkih, kot so dihalne pavze ali premori pred začetki novih kitic, se odražajo v medsebojno podobnih vrednostih značilnic MFCC. To je tudi razlog, da metoda najde toliko pravilnih mej segmentov v posnetkih.

3.3 MSAF-HPCP-SCluster

Metoda, predstavljena v [McFee and Ellis, 2014a], uporablja tehnike s področja spektralne teorije grafov za hierarhično segmentacijo na podlagi matrike ponovitev, izračunane na podlagi značilnic MFCC in HPCP. Rezultati metode so precej enakovredni med različnimi sestavi, vrednost mere F1 pa se giblje okoli 0,5 za vse sestave. Rezultati metode odražajo tudi zelo uravnotežene vrednosti priklica in natančnosti za neinstrumentalno glasbo, po drugi strani metoda precej nadsegmentira instrumentalno glasbo.

Trije poglavitni razlogi za takšno obnašanje metode so:

1. metoda ni občutljiva zgolj na spremembe v tempu le, če so te majhne, saj ve-

čje spremembe v tempu vplivajo na spektralno gručenje in s tem na končno segmentacijo;

2. metoda je občutljiva na drsenje višine tonov;

3. visoka nadsegmentacija instrumentalne glasbe nakazuje ali na neprimerno ravnovesje med globalno in lokalno povezljivostjo v grafu segmentacije ali pa na neprimerno izbiro končnih meja segmentov iz hierarhične segmentacije.

Med vsemi izbranimi metodami metoda MSAF-HPCP-SCluester za noben sestav ni najuspešnejša, je pa druga najuspešnejša metoda (med izbranimi in tretja med vsemi) na celotni zbirki glede na mero F1.

3.4 MSAF-MFCC-SF

Splošen pristop k segmentaciji časovnih vrst [Serrà et al., 2012] temelji na segmentacijskih značilnicah, izračunanih iz filtrirane matrike časovnih zakasnitev. Metoda je po uspešnosti primerljiva z metodo MSAF-HPCP-SCluster in na celotni zbirki doseže vrednost mere F1 0,47. Nad- in podsegmentacija nista prisotni v segmentaciji neinstrumentalne glasbe, sta pa bolj izraziti pri instrumentalni glasbi.

Metoda je zasnovana tako, da ni občutljiva na majhne spremembe v tempu, ni pa uspešna pri večjih spremembah, ki se pojavljajo v posnetkih petja. To vpliva na dva koraka metode:

1. na uporabo blokov značilnic (koordinate časovne zakasnitve), ki za izračun matrike časovne zakasnitve vrne slabo oceno podobnosti in posledično slabo segmentacijo, saj se lahko tempo med posameznimi ponovitvami precej razlikuje;

2. na izračun značilnic podobnosti, saj so značilnice in posledično krivulja novitet zamazane v času.

Metoda je prav tako občutljiva na drsenje višine tonov, saj za izračun matrike časovnih zakasnitev uporablja Evklidsko normo.

3.5 Diskusija

Z analiziranjem 10 % pesmi, kjer so se metode odrezale najslabše, smo ugotovili, da v tem naboru ni veliko skupnih pesmi za vse metode. Metodi, ki uporabljata značilnice MFCC (to sta MSAF-MFCC-Foote in MSAF-MFCC-SF), imata v naboru najslabših 10 % (kar predstavlja 20 pesmi) skupne samo 4 pesmi. Par metod MSAF-MFCC-Foote in MSAF-HPCP-SCluster imata v tem naboru samo 3 skupne pesmi, par metod MSAF-MFCC-SF in MSAF-HPCP-SCluster pa zgolj 6 pesmi. Vse tri metode pa imajo med najslabšimi 10 % rezultatov samo eno skupno pesem. To nakazuje na dejstvo, da razlog, zakaj metode odpovedo, ni nekaj *težkih* pesmi, ampak da vsaki metodi spodleti zaradi drugačnih razlogov. Nekaj teh razlogov smo predstavili v predhodnih podpoglavjih.

Glede na pomnjkljivosti obstoječih metod smo se odločili, da bomo pri zasnovi lastne metode upoštevali tudi specifike ljudske glasbe. Naša metoda bo tako upoštevala naslednje:

1. odpornost na razlike v tempu in izračunu podobnosti med posameznimi segmenti;
2. odpornost na drsenje višine tonov pri izračunu podobnosti med posameznimi segmenti;
3. odpornost na šum in napake izvajalcev, do katerih prihaja v pesmih;
4. pesmi so sestavljene iz ponovitev enega melodičnega ali harmoničnega vzorca;
5. poudarek je na neinstrumentalni glasbi, ki predstavlja večji izziv za trenutne segmentacijske metode kot instrumentalna glasba.

4

Osnovni pojmi in uporabljene metode

Everything, however complicated - breaking waves, migrating birds, and tropical forests - is made of atoms and obeys the equations of quantum physics. But even if those equations could be solved, they wouldn't offer the enlightenment that scientists seek. Each science has its own autonomous concepts and laws.

– Martin Rees

V tem poglavju predstavimo osnovne pojme in metode, ki jih uporabljamo pri izdelavi lastne segmentacijske metode.

4.1 Kromatične značilnice

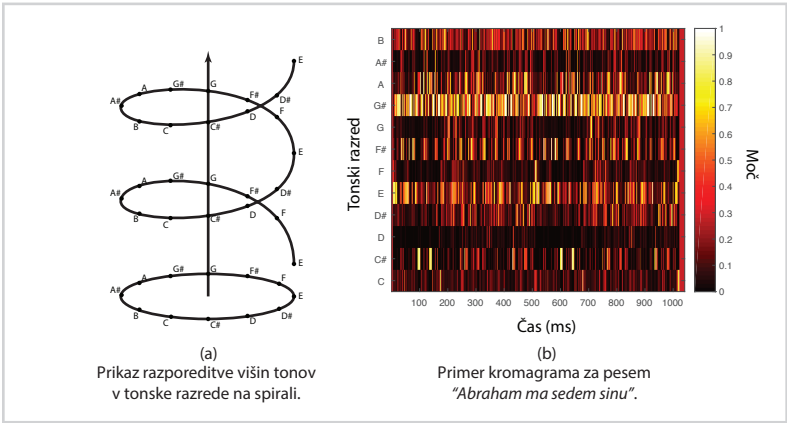
Velik del zahodne glasbe temelji na tonskem sistemu, ki razporeja zvoke glede na razmerja med višinami tonov v soodvisne prostorske in časovne strukture, kot so akordi, ključi, melodije in motivi. Višino tona (angl. pitch) zaznavamo ciklično med različnimi oktavami. Toni se med posameznimi oktavami ponavljajo in jih združujemo v t. i. tonske oz. kromatične razrede (angl. pitch classes ali chroma classes). Ravno predstavljene ciklične ponovitve tonov nam omogočajo razdelitev zvoka glede na višino tona in tonski razred [Goto, 2006]. Vseh tonskih razredov je 12 in so: C, C#, D, D#, E, F, F#, G, G#, A, A#, H, med seboj pa so razmaknjeni za en polton. Kako se toni združujejo v tonske razrede glede na višino, je prikazano na sliki 4.1 (a), kjer so višine tonov nanizane na spiralo, kot na spirali pa določa tonski razred. Seveda lahko tonske razrede diskretiziramo podrobneje kot zgolj na en polton z uporabo centov. Velja, da je en polton enak 100 centov.

Tonskim lastnostim pravimo tudi kromatične lastnosti zvoka, ki jih lahko modeliramo s kromatičnimi značilnicami. Primer takšnih značilnic so kromatični vektorji in njihove različne izpeljanke, kot so značilnice CENS in značilnice HPCP. Nekaj izbranih značilnic, ki jih uporabimo ali omenjamo v našem pristopu, je predstavljenih v nadaljevanju.

4.1.1 Kromatični vektorji

Kromatični vektor (angl. chroma vector) je predstavitev zvoka v obliki $n \times 12$ dimenzionalnega vektorja, kjer posamezna vrednost v vektorju predstavlja vsoto frekvenčnih komponent posameznega tonskega razreda v opazovanem okvirju zvočnega posnetka. Pri osnovnem pristopu za izračun kromatičnih značilnic zvočni signal razdelimo v 88 tonskih pasov glede na vsebovano frekvenco. Srednjo frekvenco tonskega pasu lahko izračunamo kot:

$$f_{c,k} = f_{\min} \cdot 2^{\frac{k}{\beta}}, \tag{4.1}$$



(a) Prikaz razporeditve višin tonov v tonske razrede na spirali.

(b) Primer kromagrama za pesem *"Abraham ma sedem sinu"*.

Slika 4.1: Sliki prikazujeta razporeditev višin tonov v tonske razrede in kromagram.

kjer je $f_{\min}$ minimalna analizirana frekvenca, k je indeks filtra iz $[0, (\beta \times Z) - 1]$, β predstavlja število tonskih razredov na oktavo in Z predstavlja število oktav.

Za vsak takšen pas izračunamo kratkočasovno povprečno energijo (angl. short-time mean-square power - STMSP) in seštejemo vse vrednosti STMSP, ki ustrezajo istemu tonskemu razredu [Müller, 2007]. Na ta način dobimo 12-dimenzionalni kromatični vektor za posamezni okvir signala:

$$C_b = \sum_{z=0}^{Z-1} \|X_{lf}(b + z\beta)\|, \tag{4.2}$$

kjer je X_{lf} frekvenčni spekter, z oktavni index na intervalu $[0,\ Z-1]$, Z število oktav, b indeks tonskega (kromatičnega) razreda iz intervala $[0,\ \beta - 1]$ in β število tonskih razredov na oktavo.

Primer takšnega vektorja predstavlja posamični stolpec celotne strukture, prikazane na sliki 4.1 (b). Več posameznih kromatičnih vektorjev lahko združimo v enotno strukturo, imenovano kromagram, ki je prikazana na sliki 4.1 (b).

Z določenimi modifikacijami pri izračunu in pri drugačni razporeditvi frekvenčnih pasov pa lahko takšne značilnice na različne načine modelirajo kromatične lastnosti zvočnega posnetka. V nadaljevanju predstavimo dva takšna primera.

4.1.2 Značilnice CENS

Značilnice CENS (angl. chroma energy normalized statistics) so predstavljene v delu [Müller, 2007]. Značilnice dodajo nov nivo abstrakcije z upoštevanjem kratkočasovne statistike nad porazdelitvijo energije znotraj posameznih tonskih pasov. Značilnice so skalabilne in robustne ter močno korelirane s kratkočasovno vsebino zvočnega signala. Značilnice obstajajo v različnih izvedenkah, kot je na primer izvedenka, ki podrobneje modelira osnovno frekvenco v posnetkih - t. i. F0 ojačane značilnice CENS (angl. F0 enhanced CENS features).

4.1.3 Značilnice HPCP

Značilnice HPCP (angl. harmonic pitch class profiles) so izračunane na podlagi profilnice tonskega razreda (angl. pitch class profile) [Gómez, 2006]. Značilnice modelirajo tonaliteto z merjenjem relativne intenzitete za vsak kromatični razred znotraj izbranega časovnega okvirja. Značilnice so odporne na šum okolja in zvoke tolkal. So neodvisne od barve zvoka in instrumentacije, neodvisne pa so tudi od glasnosti in dinamike.

Značilnice HPCP so izračunane z uporabo magnitude vrhov spektra znotraj določenega frekvenčnega pasu, ki predstavljajo najbolj reprezentativne frekvence, ki nosijo harmonične lastnosti. Pri izračunu se uporabijo tudi uteži, s katerimi kompenziramo razlike v uglasitvi in neharmoničnosti. Tako izračunan vektor je na koncu normaliziran, s čimer zavržemo informacijo o energiji. Značilnice HPCP predstavimo kot:

$$h_n = \sum_{i=1}^{\text{nPeaks}} w(n, f_i) \cdot a_i^2;\ n = 1 \ldots \text{size}, \tag{4.3}$$

kjer sta a_i in f_i linearna magnituda in frekvenca i-tega lokalnega maksimuma spektra, nPeaks je število lokalnih maksimumov spektra, ki jih upoštevamo, n je kromatični razred značilnice h, size je ločljivost značilnice h (število razredov: 12, 24 ...) in $w(n, f_i)$ je utež frekvence f_i pri upoštevanju za n-ti razred značilnice h.

Prav te značilnice smo sklenili uporabiti tudi v našem pristopu, saj se je po testiranju izkazalo, da za naravo materiala v predstavljeni glasbeni zbirki najbolje modelirajo lastnosti v zvočnem posnetku. Kot vse kromatične značilnice tudi za značilnice HPCP velja,

da lahko njihovo ločljivost prilagodimo. Za naše potrebe smo izbrali 24-dimenzionalne značilnice HPCP.

4.2 Podobnostne strukture

Za prikaz podobnosti med posameznimi elementi neke množice obstaja veliko različnih struktur, kot so grafi, mreže, zemljevidi ipd. Za prikaz podobnosti znotraj neke časovne vrste pa se najpogosteje uporabljata dva načina predstavitve. Prvi način predstavlja samopodobnostna matrika (angl. self-similarity matrix), drugi način pa njena sorodna predstavitev - matrika časovnih zamikov (angl. time-lag matrix). Oba načina prikaza sta v nadaljevanju nekoliko podrobneje predstavljena.

4.2.1 Samopodobnostna matrika

Samopodobnostna matrika je grafična predstavitev podobnosti med posameznimi deli zaporedja. V časovnem signalu je lahko to podobnost med pari časovnih okvirjev signala. Primer takšne matrike je prikazan na sliki 4.2 (a), v matriki so označeni tudi ponavljajoči vzorci. Na obeh oseh dvodimenzionalne predstavitve je čas. Vrednost v posamezni celici matrike pa predstavlja stopnjo podobnosti med časovnima okvirjema signala glede na izbrano mero podobnosti. Podobni deli se tako v matriki izražajo kot poudarjene diagonalne sledi. Glavna diagonala predstavlja osnovni signal, vzporedne sledi pa med sabo povezujejo dele s podobnimi lastnostmi glede na izbrano mero podobnosti. Primeri takšnih mer podobnosti za primera vektorjev $x = (x_1, \ldots, x_n)$ in $y = (y_1, \ldots, y_n)$ so:

- *Evklidska razdalja* - predstavlja razdaljo med dvema vektorjema x in y, definirano kot:

$$d_{\text{evk}} = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2} \tag{4.4}$$

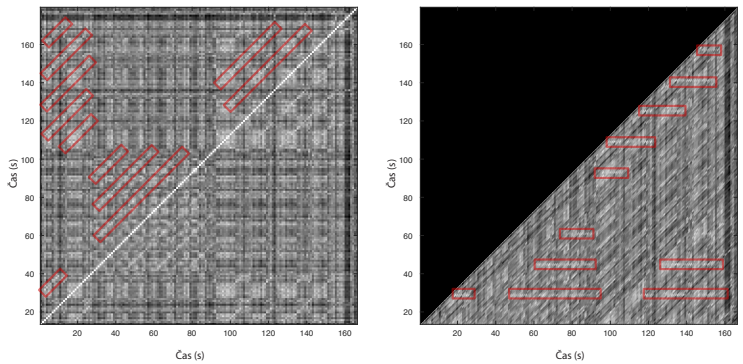
- *Kosinusna razdalja* - predstavlja razdaljo med dvema vektorjema x in y, definirano kot:

$$d_{\cos} = 1 - \frac{x_1 y_1 + \ldots + x_n y_n}{\sqrt{|x_1|^2 + \ldots + |x_n|^2}\ \sqrt{|y_1|^2 + \ldots + |y_n|^2}} \tag{4.5}$$

- *Korelacijska razdalja* - predstavlja razdaljo med dvema vektorjema x in y, definirano kot:

$$d_{\text{corr}} = 1 - \frac{(x - \bar{x})\,(y - \bar{y})}{||x - \bar{x}||\;||y - \bar{y}||} \tag{4.6}$$

Samopodobnostne matrike predstavljajo posebno obliko bolj znane statistične metode izrisa ponovitev (angl. recurrence plot), ki je namenjena odkrivanju ponavljajočih dogodkov v časovnih vrstah [Eckmann et al., 1987], za katere velja, da sta si izbrana vektorja x in y podobna glede na izbrano mero podobnosti. Na področje MIR je samopodobnostne matrike vpeljal Foote z delom [Foote, 1999], kjer za mero podobnosti vzame značilnice MFCC, ki modelirajo barvo zvoka.



(a) Samopodobnostna matrika. (b) Matrika časovnih zamikov

Slika 4.2: Sliki prikazujeta samopodobnostno matriko in matriko časovnih zamikov.

4.2.2 *Matrika časovnih zamikov*

Samopodobnostni matriki enakovredna predstavitev podobnosti v nekem signalu je matrika časovnih zamikov. Pri tej obliki predstavitve podobnosti med posameznimi deli signala se ponovitve v matriki ponovitev pojavljajo vzporedno z vodoravno osjo in ne več vzporedno z diagonalo. Prednost takšne predstavitve je lažje odkrivanje sledi v strukturi, saj so le-te vodoravne. Matrike časovnih zamikov je na področje MIR vpeljal Goto, ki njihovo uporabo predstavi v delu [Goto, 2003]. Primer matrike časovnih zamikov je prikazan na sliki 4.2 (b).

4.3 Dinamično ukrivljanje časa

Dinamično ukrivljanje časa (angl. dynamic time warping - DTW) je prilagojena oblika dinamičnega programiranja. Metoda je namenjena odkrivanju optimalne poravnave med dvema (časovno odvisnima) zaporedjema in je bila primarno uporabljena v sistemih za avtomatsko transkripcijo govornih posnetkov. DTW je primerna metoda za premoščanje časovnih transformacij v signalu in razlik v hitrosti pri časovno odvisnih zaporedjih, kar predstavlja najpogostejšo uporabo na področju MIR.

Kot vhod metoda vzame dve zaporedji, predstavljeni z vektorjema x in y:

$$x = (x_1, x_2, \dots, x_N), y = (y_1, y_2, \dots, y_M);\ M, N \in \mathbb{N}. \tag{4.7}$$

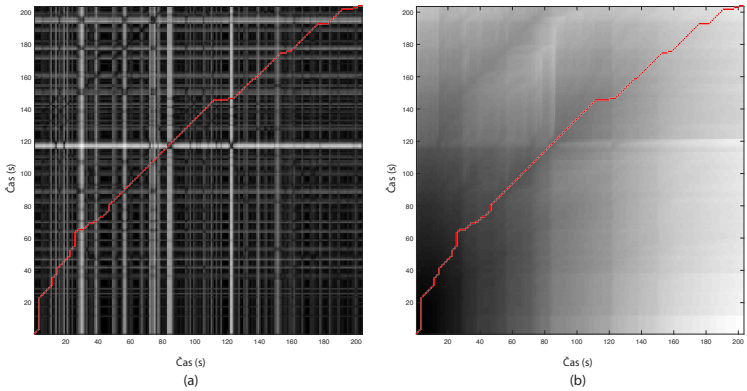
Iz podanih vektorjev naračunamo samopodobnostno matriko C, katere vrednosti C_{ij} predstavljajo razdaljo med elementi vhodnih vektorjev i in j (takšna matrika je prikazana na sliki 4.3 (a)). Za razdaljo med posameznimi vektorji lahko izberemo različne mere razdalje, so kot na primer evklidska razdalja, kosinusna razdalja, korelacijska razdalja ipd. Cilj je poiskati poravnavo med x in y z najmanjšo razdaljo in posledično največjo podobnostjo. Takšna pot poteka po vrednostih lokalnih minimumov s čim nižjimi razdaljami v samopodobnostni matriki C. Rezultat je ukrivljena pot skozi matriko kumulativne funkcije razdalje (glej sliko 4.3 (b)), ki je predstavljena z zaporedjem $p = (p_1 \dots p_L);\ p_l = (n_l, m_l) \in [1 : N] \times [1 : M];\ l \in [1 : L]$ in ustreza:

- *robnim pogojem*: $p_1 = (1, 1)$ in $p_L = (N, M)$;
- *monotonosti*: $n_1 \leq n_2 \leq \dots \leq n_L$ in $m_1 \leq m_2 \leq \dots \leq m_L$;
- *pogoju koraka*: $p_{l+1} - p_l \in \{(1, 0), (0, 1), (1, 1)\}$ za $l \in [1 : L - 1]$.

Kumulativno razdaljo $d_p(x, y)$ ukrivljene poti p med x in y glede na funkcijo razdalje d izračunamo kot:

$$d_p(x, y) = \sum_{l=1}^{L} d(x_n, y_m). \tag{4.8}$$

Optimalna ukrivljena pot p^* je tista, ki ima izmed vseh ukrivljenih poti minimalno



Slika 4.3: Sliki prikazujeta podobnostno matriko (a) in kumulativno matriko (b) z označeno optimalno ukrivljeno potjo (rdeče).

kumulativno razdaljo. Imenujemo jo tudi DTW in je definirana kot:

$$DTW = d_{p^*}(x, y) = min\{d_p(x, y) \mid p \text{ je } (N, M) - \text{ukrivljena pot}\} \tag{4.9}$$

Pri izračunu DTW večinoma upoštevamo različne omejitve pri iskanju ukrivljene poti. S takšnimi omejitvami lahko pohitrimo sam izračun, kot tudi preprečimo nezaželene načine poravnave z globalnim usmerjanjem ukrivljene poti znotraj nekega področja. Dve dobro znani področji, na kateri lahko omejimo iskanje, sta Sakoe-Chibajev pas in Itakurin paralelogram [Müller, 2007].

Za učinkovito iskanje optimalne ukrivljene poti p^* lahko uporabimo dinamično programiranje, pri čemer sestavimo matriko D, imenovano kumulativna matrika razdalj, katere primer je na sliki 4.3 (b), in jo definiramo kot:

$$D(n, m) = DTW(x(1 : n), y(1 : m)). \tag{4.10}$$

Takšen pristop je priročen za ugotavljanje podobnosti v zvočnem posnetku kljub spremembam v tempu. Na področju MIR je uporabo dinamičnega ukrivljanja časa dobro predstavil Müller v delu [Müller, 2007].

4.4 Skriti markovski model

Skriti markovski model (angl. hidden Markov model - HMM) je statistični model, ki temelji na predpostavki, da lahko modelirani sistem obravnavamo kot markovski proces s skritimi stanji. V večini primerov se želimo modela naučiti iz podatkov in s tem poiskati statistični model, ki dobro modelira strukturo podatkov. Vsa stanja HMM-ja imajo verjetnostno porazdelitev preko vseh izhodnih simbolov, parametri sistema pa so verjetnosti prehodov med posameznimi stanji in povprečja ter variance za posamezna stanja.

Formalno diskretni HMM opišemo s peterko (S, V, Π, A, B), kjer je $S = \{s_1, \dots, s_N\}$ množica N stanj sistema, $V = \{v_1, \dots, v_M\}$ množica M izhodnih simbolov, $\Pi = \{\pi_1, \dots, \pi_N\}$ porazdelitev verjetnosti začetnih stanj, $A = \{a_{ij}\}$ matrika verjetnosti prehodov med stanji modela in $B = \{b_{i,v_k}\}$ matrika verjetnosti oddajanja simbolov sistema. Z množico $\Lambda = (\Pi, A, B)$ predstavimo parametre sistema, ki predstavljajo:

- π_i – predstavlja verjetnost, da sistem prične v stanju i,
- a_{ij} – predstavlja verjetnost prehoda iz stanja i v stanje j in
- b_{i,v_k} – predstavlja verjetnost, da sistem v stanju i odda simbol v_k.

Veljati mora tudi:

$$\sum_{i=1}^{N} \pi_i = 1,$$

$$\sum_{j=1}^{N} a_{ij} = 1;\ i \in [1,\ N],$$

$$\sum_{k=1}^{M} b_{i,v_k} = 1;\ i \in [1,\ N].$$

Podrobnejši opis modela, procesa učenja parametrov in ugotavljanja najverjetnejšega stanja za posamezni okvir zvočnega signala pri razpoznavi govora predstavi Rabiner v delih [Rabiner, 1989; Rabiner and Juang, 1993].

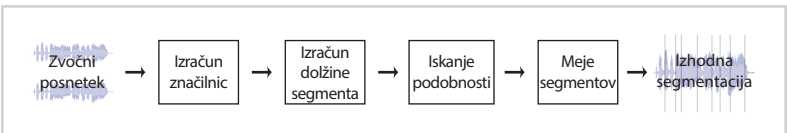
5

Metoda za segmentacijo ljudskih pesmi

[The theory of universal gravitation] is not cast-iron. No theory is, and there is always room for improvement. Isn't that so? Science is constructed out of approximations that gradually approach the truth. . . Well, that means all theories are subject to constant testing and modification, doesn't it? And if it eventually turns out that they're not quite close enough to the truth, they need to be replaced by something that's closer. Right?

– Isaac Asimov

Segmentacijski pristop, razvit v okviru disertacije, vzame za vhod zvočni posnetek ljudske pesmi, ga procesira v več fazah in kot končni rezultat vrne seznam mej med posameznimi segmenti. Z razvito metodo želimo poiskati meje med posameznimi segmenti pesmi, pri čemer želimo, da so posamezni deli med seboj čimbolj melodično podobni, hkrati pa približno enake dolžine. Prav tako želimo, da je metoda robustna na spremembe v tempu, na spremembe v intonaciji, do katerih prihaja zaradi drsenja višine tonov, kot tudi na lokalne napake, do katerih prihaja, kadar je napačno zapet kakšen ton. Posamezni koraki pristopa so prikazani na sliki 5.1. Vhodni signal najprej pretvorimo v kromatično predstavitev, ki jo nadalje uporabimo za izračun krivulj oddaljenosti (angl. distance curves), neobčutljivih na spremembe v tempu in drsenje v višini tonov. Te krivulje nakazujejo stopnjo podobnosti med posameznimi deli signala. Ker predpostavimo, da pesem vsebuje ponovitve enega melodičnega vzorca, z uporabo krivulj oddaljenosti izračunamo povprečno dolžino takšnega vzorca. Za izračun mej med posameznimi segmenti uporabimo verjetnostni pristop, kjer maksimiziramo verjetnost postavitve posameznih mej med segmenti glede na zastavljene omejitve (omejitev dopustnega drsenja višine tonov, omejitev v sprejemljivi stopnji melodične podobnosti ipd.). Glavna izhoda metode sta ocena drsenja intonacije pri petju in meje med posameznimi segmenti pesmi.



Slika 5.1: Metoda in njeni posamezni koraki.

Segmentacijo predstavimo z množico časov Ψ, ki jo definiramo kot

$$\Psi = \{t_1, t_2, \ldots, t_G\}, \tag{5.1}$$

kjer t_i predstavlja čas začetka i-tega segmenta, G pa število vseh segmentov v posnetku. V nadaljevanju so predstavljeni posamezni koraki metode.

5.1 Predstavitev z značilnicami

Vhodni zvočni signal s preberemo iz zvočne datoteke, ga povprečimo v en sam kanal in normaliziramo tako, da vrednosti delimo z maksimalno absolutno vrednostjo v si-

gnalu, s čimer izenačimo delovanje metode med glasnejšimi in tišjimi posnetki. Za nadaljnjo uporabo signal s, ki je večinoma vzorčen z vzorčevalno frekvenco 44.100 Hz, prevzorčimo z vzorčevalno frekvenco 22.050 Hz. Prevzorčenje izvedemo zaradi hitrejših izračunov nadaljnjih korakov metode, kar si lahko dopustimo, ker frekvence nad 11.050 Hz v signalu ne nosijo koristnih informacij.

5.1.1 Izbira značilnic

Za predstavitev vsebine zvočnega signala uporabimo kromatične značilnice, saj le-te najbolje zajamejo melodične in harmonične podobnosti med posameznimi ponovitvami znotraj pesmi. Ker obstaja veliko različnih kromatičnih značilnic, smo na majhni zbirki pesmi preizkusili, kako dobro zajemajo značilnosti signala nekatere izmed najpogosteje uporabljenih značilnic. Medsebojno smo primerjali za F0 ojačane značilnice CENS [Müller, 2007] in značilnice HPCP [Gómez, 2006].

Med primerjavo obeh vrst značilnic smo prišli do ugotovitve, da značilnice HPCP bolje modelirajo harmonično vsebino glasbe v izbranih posnetkih naše glasbene zbirke. Razlog je v tem, da so za F0 ojačane značilnice CENS bolj primerne za solo petje, saj ojačajo prevladujočo frekvenco, v naši zbirki pa imamo glasbo različnih glasbenih sestavov.

Drugi aspekt značilnic je njihova ločljivost, ki predstavlja en polton oz. en tonski razred. Izbrane značilnice lahko modelirajo signal z različnim številom kromatičnih razredov. V osnovi imajo značilnice 12 kromatičnih razredov, kjer vsak izmed razredov predstavlja en polton. Po zgledu prispevka [Serrà et al., 2008] smo medsebojno primerjali tudi uporabo kromatičnih značilnic pri različnih ločljivostih. Prišli smo do enakih ugotovitev kot avtorji omenjenega prispevka: večja ločljivost značilnic bolje modelira kromatične značilnosti signala. Razlog, da se 24-dimenzionalne značilnice obnesejo bolje od 12-dimenzionalnih je tudi v tem, da ljudsko petje nima točne intonacije in s tem bolje zajamemo vmesne višine tonov, ki bi se sicer razlili preko sosednjih poltonov.

V našem primeru smo tako uporabili značilnice s 24 razredi, kjer en polton pokriva dva razreda oz. en razred pokriva 50 centov. Posamezna kromatična značilnica je predstavljena z vektorjem realnih vrednosti dimenzije, enake številu kromatičnih razredov značilnice, in modelira kromatično vsebino kratkega dela zvočnega signala.

5.1.2 Izračun značilnic

Zaradi zgoraj navedenih dejstev smo se odločili v končni implementaciji metode uporabiti 24-dimenzionalne značilnice HPCP, kjer posamezno označimo s h. Značilnice smo računali s korakom 100 ms in dolžino okna 200 ms. Ker ne potrebujemo zelo natančne časovne ločljivosti, smo izbrali večjo dolžino okna, kar zmanjša vpliv šuma v signalu. Za izračun značilnic HPCP smo uporabili programsko knjižnico Essentia [Bogdanov et al., 2013].

Za vhodni zvočni signal s izračunamo značilnice HPCP in jih predstavimo kot zaporedje H:

$$H = (h_1, h_2, \ldots, h_{N_h}), \tag{5.2}$$

kjer h_i predstavlja 24-dimenzionalno značilnico HPCP, N_h pa predstavlja število vseh HPCP značilnic v signalu.

5.2 Računanje podobnosti

Izbira ustrezne mere podobnosti je ključnega pomena za uspešno segmentacijo. Večina trenutnih pristopov uporablja lokalne mere podobnosti, ki medsebojno primerjajo kratkočasovne značilnice, kot so kromatične značilnice ali značilnice MFCC. Posamezni kratek del signala tako primerjajo z vsemi ostalimi deli signala, kar lahko prikažemo s samopodobnostno matriko ali s strukturo časovno zamaknjenih ponovitev podobnih značilnic v signalu. Pri tem smo omejeni na primerjavo kratkočasovnih odsekov signala, nad katerimi so izračunane posamične značilnice. Izjemo predstavlja pristop [Serrà et al., 2012], kjer avtorji medsebojno primerjajo časovna zaporedja značilnic in ne zgolj posameznih vrednosti. S tem ne povzemajo samo lokalnih podobnosti med značilnicami, ampak podobnosti parov kratkih zaporedij značilnic. Večina pristopov obravnava variacije v tempu šele v kasnejših korakih segmentacije ali pa variacije v tempu v celoti zanemarijo.

Kot je razvidno iz ovrednotenja trenutno aktualnih segmentacijskih metod na zbirki ljudske glasbe, metode zaradi takšne zasnove niso sposobne upoštevati večjih variacij v tempu in vračajo zaradi tega slabše rezultate.

To je bil tudi poglavitni razlog, da smo se pri implementaciji lastne metode odločili variacije v tempu obravnavati že pri samem izračunu podobnostnih vrednosti med posameznimi deli signala. Odločili smo se za podobno rešitev, kot jo predstavijo avtorji v prispevku [Müller et al., 2009]. V prispevku avtorji predlagajo uporabo dinamičnega ukrivljanja časa (DTW) za izračun podobnosti med dvema deloma zvočnega signala pesmi. DTW, ki smo ga predstavili v predhodnem poglavju, je običajna tehnika za izmero podobnosti med dvema časovnima vrstama ali deloma časovnih vrst, ki se lahko razlikujeta v trajanju.

Naj bo $H_{t_1}^{(T)}$ podzaporedje zaporedja H, ki se začne v času t_1 in traja T sekund, in naj bo $c(h_i, h_j)$ mera razdalje med značilnicama h_i in h_j. Podzaporedje $H_{t_i}^{(T)}$ predstavlja posamezni segment, kjer se prva značilnica začne v času t_i in konča v času $t_i + T$.

5.2.1 Izbira in izračun mere podobnosti

Kot smo predstavili v prejšnjem poglavju, obstajajo različne mere razdalj, bolj ali manj primernih za uporabo z zvočnimi signali. V primeru iskanja podobnosti med kromatičnimi značilnicami se najpogosteje uporabljata kosinusna razdalja in korelacijska razdalja. Eksperimentalno smo ovrednotili, katera se obnese bolje pri uporabi s 24-razrednimi značilnicami HPCP, in prišli do enakih dognanj kot avtorji v prispevku [Serrà et al., 2008]. Bolje se obnese uporaba korelacijske razdalje, ki je definirana kot 1 minus korelacija med dvema vektorjema. Korelacijska razdalja med dvema značilnicama HPCP h_i in h_j je definirana kot:

$$c(h_i, h_j) = 1 - \frac{(h_i - \bar{h})\,(h_j - \bar{h})}{||h_i - \bar{h}||\;||h_j - \bar{h}||}\,, \tag{5.3}$$

kjer $\bar{h}$ predstavlja povprečno značilnico HPCP.

Za izračun kumulativne mere razdalj d_{ij} med dvema zaporedjema značilnic $H_i^{(L_1)}$ in $H_j^{(L_2)}$ izračunamo optimalno pot časovne ukrivljenosti med dvema zaporedjema z minimizacijo skupne razdalje kot:

$$d_{ij} = \min \frac{1}{|w|} \sum_{(m,n)\in w} c(h_m, h_n), \tag{5.4}$$

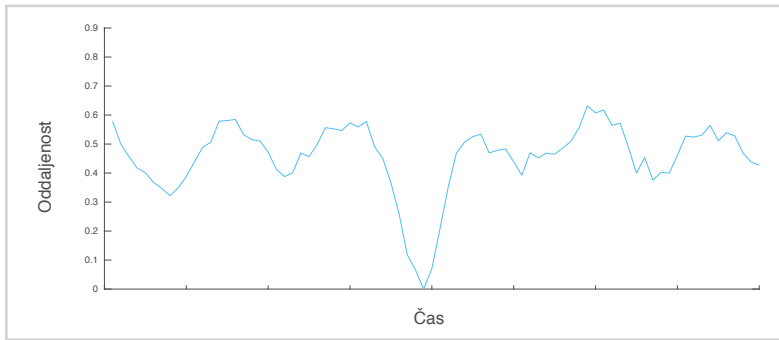
kjer w predstavlja pot časovne ukrivljenosti s pričetkom na začetku obeh zaporedij (i, j) in koncem ob koncu zaporedij $(i + L_1 - 1, j + L_2 - 1)$ in velja, da $h_m \in H_i^{(L)}$ in $h_n \in H_j^{(L)}$. Problem lahko rešimo z dinamičnim programiranjem s časovno kompleksnostjo $O(n^2)$.

Za izračun DTW obstajajo različni pristopi. Poleg osnovnega pristopa, predstavljenega v predhodnem poglavju, lahko pri izračunu optimalne poti uporabimo tudi različne omejitve glede na to, kakšne skoke dopuščamo med posameznimi podobnostnimi elementi. V našem primeru smo uporabili izvedenko DTW za uporabo s podzaporedji (angl. subsequence DTW) [Müller, 2007], kjer smo za mero razdalje med posameznimi segmenti uporabili korelacijsko razdaljo. Tako prilagojena tehnika dinamičnega ukrivljanja časa dopušča, da se izbrana optimalna pot ne zaključi s primerjavo zadnjih elementov posameznih zaporedij, ampak se lahko konča tudi pred koncem enega izmed zaporedij. Tehniko s podzaporedji izberemo zaradi tega, ker nimamo zagotovila, da izbrana podzaporedja predstaljajo enake dele ponovljenih segmentov.

Pri izračunu podobnosti odstranimo okvirje signala s tišino, saj med seboj ne želimo primerjati delov tišine, ki lahko predstavljajo visoko stopnjo podobnosti, ki pa ne odražajo podobnosti v petju. Enako velja za različno dolga območja signala, ki predstavljajo dihalne pavze: ne želimo, da različno dolge dihalne pavze znižajo stopnjo podobnosti med ponovljenimi segmenti.

5.2.2 *Izračun krivulj oddaljenosti*

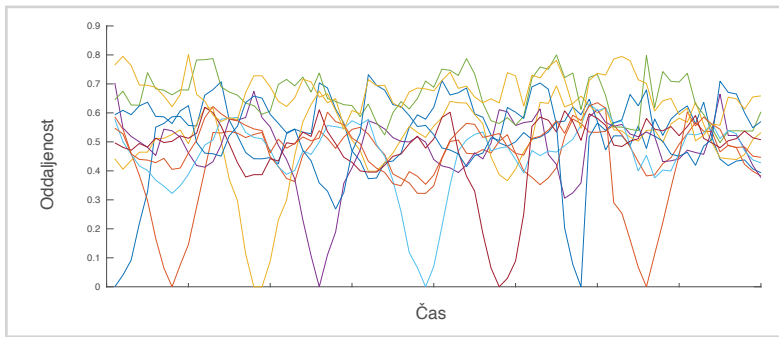
Izračun podobnosti med posameznimi okvirji lahko predstavimo s t. i. krivuljo oddaljenosti, kjer posamezna vrednost krivulje predstavlja vrednost mere oddaljenosti med pari okvirjev primerjanih delov posnetka. Iz oblike krivulje lahko ugotovimo, kje se v signalu stvari ponavljajo - kjer se v krivuljah nahajajo nižje vrednosti. *Krivulja oddaljenosti* $D_i = (d_{i1}, d_{i2}, \ldots, d_{iN})$ predstavlja razdalje med segmentom pesmi $H_i^{(L)}$ in vsemi ostalimi segmenti dolžine L v pesmi, ki se pričnejo vsako sekundo signala. Primer takšne krivulje je predstavljen na sliki 5.2. Takšna krivulja prikazuje podobnosti med izbranim delom in celotnim zvočnim signalom. Lokalni minimumi ponazarjajo ponovitve.



Slika 5.2: Primer krivulje oddaljenosti.

5.2.3 *Izračun povprečne krivulje oddaljenosti*

Posamezna krivulja oddaljenosti ne zajame nujno podobnosti skozi celotni signal. Zato izračunamo več krivulj, jih poravnamo in povprečimo ter s tem izračunamo povprečno krivuljo oddaljenosti. Krivulje oddaljenosti izračunamo za J delov signala, katerih začetki v $t_i \in T_s$ so enakomerno razporejeni skozi celotno dolžino signala z naključnimi odstopanji na vsakih $\frac{l}{2} = 10$ sekund, kjer je l povprečna dolžina kitice - 20 sekund. Z izbrano razporeditvijo zajamemo podobnosti znotraj večine ponovljenih segmentov, saj $\frac{l}{2}$ predstavlja polovico povprečne dolžine kitice. Tako izračunana povprečna krivulja oddaljenosti bolje zajame, kje v signalu se nahajajo ponovitve. Primer krivulj oddaljenosti za tako izbrane lokacije v signalu prikazuje slika 5.3.



Slika 5.3: Primer krivulj oddaljenosti.

5.2.4 Dolžina reprezentativnega segmenta

Problem glede predvidene dolžine reprezentativnega segmenta L_r rešimo tako, da za L_r izberemo neko primerno vrednost, v našem primeru $\frac{l}{2} = 10$ sekund. To je dovolj dolg odsek, da ima izračunana podobnost z uporabo DTW smisel, hkrati pa je 10 sekund približno polovica dolžine povprečne kitice in je posledično malo verjetno, da presega dolžino celotne kitice v pesmi.

5.2.5 Upoštevanje drsenja višine tonov

Nenazadnje moramo upoštevati tudi problem drsenja višine tonov, do katerega pride, kadar se intonacija izvajalcev spremeni navzgor ali navzdol na različnih mestih v pesmih. Do tega problema pride največkrat pri solo- in večglasnem petju, kjer izvajalci med izvedbo pesmi spremenijo intonacijo, relativne razdalje med posameznimi notami pa ostanejo enake. V instrumentalni glasbi in pri mešanih sestavih do drsenja višine tonov prihaja redkeje, saj izvajalci z instrumenti praviloma držijo intonacijo skozi celotno pesem.

Drsenje višine tonov negativno vpliva na izračun krivulje oddaljenosti, saj za izračun oddaljenosti uporabljamo kromatične značilnice, ki povzemajo prisotnost višin tonov v posameznem delu signala. V kolikor sta dva melodično podobna dela signala zamaknjena v višini tona, podobnost med njima ne bo takšna, kot bi pričakovali.

Problem rešimo tako, da na vsaki izbrani lokaciji $t_i \in T_s$ izračunamo množico krivulj oddaljenosti D_i^p, ki predstavljajo razdaljo med zaporedjem značilnic HPCP celotne pesmi in krožno zamaknjenim zaporedjem značilnic HPCP segmenta $H_i^{(L_r)}$ za p razredov.

Krožni zamik značilnice HPCP namreč predstavlja ravno spremembo intonacije. Zamik v eno smer predstavlja spremembo intonacije navzgor, v drugo pa spremembo intonacije navzdol. Ker uporabljamo 24-dimenzionalne značilnice, en razred predstavlja 50 centov, krožni zamik za p razredov je torej enak učinku spremembe višine tona za $p \cdot 50$ centov.

Podoben pristop so uporabili tudi avtorji v prispevku [Müller et al., 2009], kjer se pri izračunu podobnosti izbira optimalnega krožnega zamika kromatične predstavitve iz-

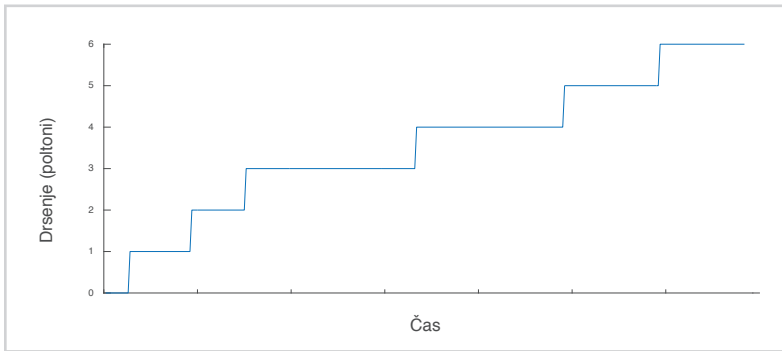
vede za vsak izračun DTW. Takšen pristop pa ni povsem realističen, saj ne upošteva dejstva, da se intonacija skozi čas ne spreminja zelo pogosto, ampak postopno preko celotne izvedbe. Ob pregledu glasbene zbirke se izkaže, da se intonacija tipično spremeni enkrat do dvakrat na kitico.

5.2.6 Omejitev drsenja višine tonov

Namesto stalnega popravljanja intonacije predlagamo omejitev drsenja višine tonov. Zaporedje izbranih vrednosti zamikov višine tonov $(p_1, p_2, \dots, p_N)$ določimo z minimizacijo razdalj preko vseh krivulj oddaljenosti D_i^p, pri čemer postavimo ceno za vsako spremembo višine tonov in s tem omejimo število sprememb $\delta(p_j - p_{j-1})$. S takšno zasnovo želimo uravnovesiti izbiro zamikov višine tonov med neomejeno minimizacijo razdalj, uporabljeno v [Müller et al., 2009], in postopnim drsenjem skozi čas:

$$\min_{p_j \in [-\zeta \dots \zeta]} \sum_{j=1}^{N} d_j^{p_j} + C_p \delta(p_j - p_{j-1}). \tag{5.5}$$

ζ predstavlja maksimalno dovoljeno tonsko drsenje, C_p pa predstavlja ceno spremembe višine tonov. Zamikanje višine tonov dopuščamo s korakom 50 centov, kar omogoča natančno spremembo intonacije. Optimizacijo izvedemo z dinamičnim programiranjem, kot rezultat pa dobimo zaporedje vrednosti zamikov višin tonov, ki se postopoma spreminjajo skozi čas. Prikaz primera drsenja višine tonov navzgor je na sliki 5.4.

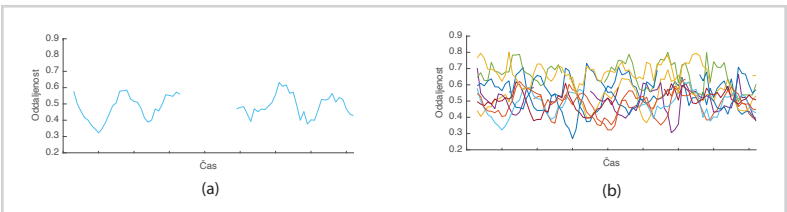


Slika 5.4: Primer drsenja višine tonov v pesmi skozi čas.

–

Rezultat opisanega postopka je množica krivulj oddaljenosti $D_i^{'} = (d_1^{p_1}, d_2^{p_2}, \ldots, d_N^{p_N})$, ki opisuje razdalje med segmenti s pričetki v $t_i \in T_{sel}$ in celotno pesmijo, ki upoštevajo spremembe v tempu in intonaciji. Takšna množica krivulj je prikazana na sliki 5.3. Ker so izbrani segmenti t_i približno enakomerno porazdeljeni po celotnem signalu, krivulje zajemajo podobnosti skozi celoten signal.

V posamezni krivulji se globalni minimum nahaja na mestu, kjer izbrani segment primerjamo ravno s tistim delom celotnega signala pesmi, kjer se ta segment nahaja. Zato je ta minimum navadno bolj izrazit od drugih. Ker ne želimo modelirati samopodobnosti izbranega segmenta, saj ti deli ne predstavljajo uporabne informacije, pred nadaljevanjem te dele iz krivulj oddaljenosti odstranimo. Rezultat za posamezno krivuljo je prikazan na sliki 5.5 (a), za množico krivulj na primeru celotne pesmi pa na sliki 5.5 (b).



Slika 5.5: Primer krivulj oddaljenosti z odstranjeno samopodobnostjo.

5.3 Izračun povprečne krivulje oddaljenosti

Predpostavko, da je pesem sestavljena iz več ponovitev segmenta (kitice), želimo izkoristiti za oceno dolžine tipičnega segmenta. Ta dolžina nam v nadaljevanju omogoča hitrejše in natančnejše iskanje mej med posameznimi segmenti.

Dolžino tipičnega segmenta ocenimo z uporabo krivulj oddaljenosti, izračunanih v predhodnem koraku. Ker so posamezne krivulje oddaljenosti izračunane s primerjanjem celotne pesmi s segmenti, izbranimi na naključnih časovnih lokacijah $t_i \in T_{sel}$, med seboj v časovni domeni niso poravnane (glej sliko 5.5 (b)). Posledično jih je potrebno medsebojno poravnati na skupno referenčno krivuljo.

5.3.1 Izbira referenčne krivulje oddaljenosti

Za referenčno krivuljo d_{ref} izmed vseh krivulj oddaljenosti izberemo tisto, ki je najbolj podobna vsem ostalim krivuljam. S tem zagotovimo, da se izbrana krivulja bistveno ne razlikuje od ostalih in predstavlja najbolj tipično krivuljo. Za izračun podobnosti med referenčnima krivuljama $d_i,\ d_j \in D_i^{'}$ smo izbrali križno kovarianco:

$$\mathrm{cov}(d_i, d_j, m) = \begin{cases} \sum_{n=0}^{N-m-1} (d_{i,n+m} - \bar{d}_{i,k})\,(d_{j,n}^{'} - \bar{d}_{j,k}^{'}), & m \geq 0, \\ \mathrm{cov}^{'}(d_i, d_j, -m), & m < 0, \end{cases} \tag{5.6}$$

kjer je m časovni zamik in N dolžina krivulje. Nezveznosti v krivuljah, ki so posledica odstranitve samopodobnosti, smo nadomestili s povprečnimi vrednostmi posameznih krivulj. Za d_{ref} velja:

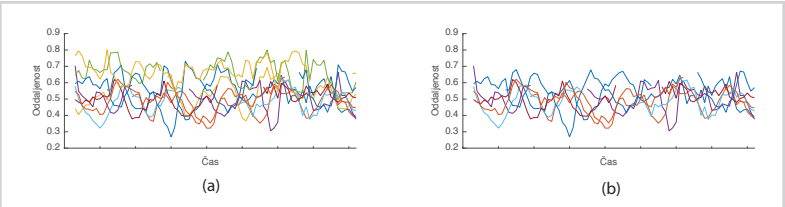
$$d_{ref} = argmax_{d_i, d_j \in D_i^{'}; m \in [-N:N]} \left\{ \mathrm{cov}(d_i, d_j, m) \right\}. \tag{5.7}$$

Za uporabo križne kovariance smo se odločili, ker deluje tudi za signale različnih dolžin, za razliko od križne korelacije, kjer je potrebno signale različnih dolžin najprej skalirati na enako dolžino.

5.3.2 Poravnava krivulj oddaljenosti

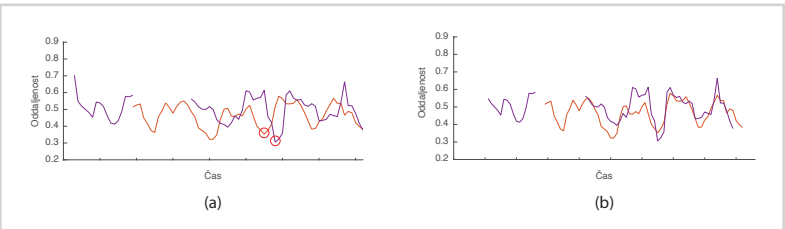
Za izračun povprečne krivulje oddaljenosti vse krivulje poravnamo z referenčno krivuljo tako, da minimum vsake krivulje poravnamo z najbližjim lokalnim minimumom referenčne krivulje.

Preden izračunamo povprečno krivuljo oddaljenosti, iz množice krivulj oddaljenosti odstranimo dele, ki navzgor značilno odstopajo od vrednosti mediane vseh krivulj, saj ti deli signala niso podobni preostalemu signalu in večinoma predstavljajo šum, ki ga ne želimo modelirati. Nadalje iz množice krivulj oddaljenosti odstranimo tiste krivulje, katerih vrednost mediane značilno odstopa navzgor glede na vrednost mediane vseh krivulj, kar je prikazano tudi na sliki 5.6 (a) pred odstranitvijo in (b) po odstranitvi.



Slika 5.6: Prikaz odstranitve odstopajočih delov krivulj in odstopajočih krivulj.

Poravnavo izvedemo tako, da najbolj reprezentativno ponovitev segmenta (minimum) posamezne krivulje poravnamo z minimumom (ponovitvijo) referenčne krivulje, ki mu je v časovni domeni najbližji. Primer poravnave je prikazan na sliki 5.7, kjer rdeča krivulja predstavlja referenčno, vijolična pa je krivulja, ki jo poravnavamo. Na sliki 5.7 (a) je prikaz krivulj pred poravnavo, na sliki 5.7 (b) pa po poravnavi. Z rdečima krogoma sta označena tudi lokalna minimuma, med katerima izvajamo poravnavo. Primer poravnav vseh krivulj, za primer izbrane pesmi, je prikazan na sliki 5.8 (a).

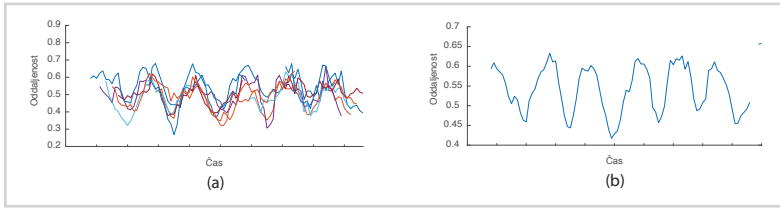


Slika 5.7: Poravnava med krivuljami oddaljenosti.

Po poravnavi krivulj na referenčno krivuljo iz njih izračunamo povprečno krivuljo oddaljenosti d_a tako, da povprečimo vse poravnane krivulje:

$$d_a = \frac{1}{M} \sum d_i^{\tau_i};\ d_i \in D_i^{'}, \tag{5.8}$$

kjer je M število krivulj oddaljenosti v množici $D_i^{'}$, τ_i pa časovni zamik pri poravnavi. Primer je prikazan na sliki 5.8 (b). Kot je vidno na sliki, so v izračunani povprečni krivulji oddaljenosti d_a posamične ponovitve segmentov bolj izražene v obliki lokalnih minimumov kot v posamičnih krivuljah. To je rezultat podobnostnih izračunov med segmenti, razporejenimi preko celotne pesmi, in ne le na posameznem segmentu.



Slika 5.8: Poravnava med krivuljami oddaljenosti.

5.4 Izračun dolžine segmenta

Za izračun dolžine segmenta uporabimo avtokorelacijo $\hat{R}(k)$, saj se podobnost dolžine segmentov v povprečni krivulji oddaljenosti odraža v obliki periodičnosti, kjer perioda predstavlja ravno dolžino tipičnega segmenta. $\hat{R}(k)$ nad povprečno krivuljo oddaljenosti d_a izračunamo:

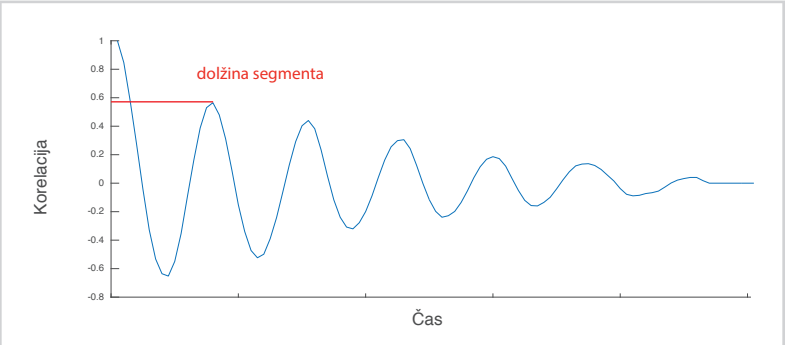
$$\hat{R}(k) = \frac{1}{(n-k)\sigma^2} \sum_{t=1}^{n-k} (d_a, t - \bar{d}_i)(d_a, t + k - \bar{d}_i), \tag{5.9}$$

kjer k predstavlja časovni zamik, N dolžino vektorja krivulje d_a, $\bar{d}_i$ povprečno vrednost vektorja krivulje d_a in σ varianco vektorja d_a.

Rezultat je avtokorelacijska funkcija, v kateri razdaljo od začetka do najvišjega vrha vzamemo za dolžino segmenta L. Primer za določitev dolžine segmenta je prikazan na sliki 5.9.

5.5 Segmentacija

Naivno bi lahko segmentacijo pridobili kar neposredno iz povprečne krivulje oddaljenosti d_a, kjer bi za meje segmentov vzeli lokalne minimume. Povprečna krivulja oddaljenosti je v času praviloma naključno zamaknjena in posledično ni nujno, da lokalni minimumi predstavljajo tudi dejanske meje med segmenti. Prav tako lahko krivulja vsebuje nepravilnosti in napake pri sami izvedbi (na primer zaradi pozabljenega in / ali ponovljenega dela kitice). Zaradi tega potrebujemo za določitev mej med



Slika 5.9: Izbrana dolžina segmenta iz avtokorelacijske krivulje.

segmenti zanesljivejši model, ki bo poleg povprečne krivulje oddaljenosti uporabil še druge informacije.

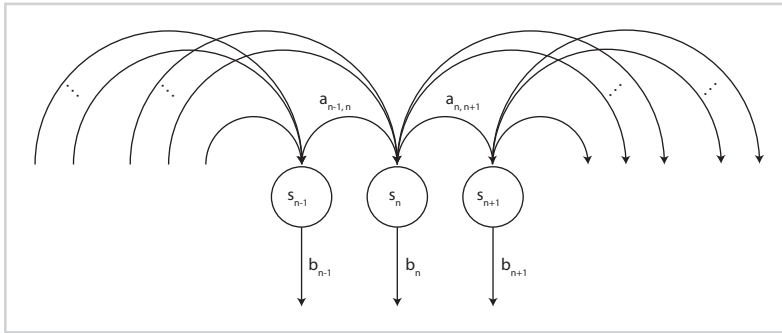
Zaradi naštetih razlogov optimalno segmentacijo izračunamo s skritim markovskim modelom, ki je definiran s peterko (S, V, Π, A, B). Model vsebuje množico stanj $S = \{s_i\};\ i \in [1, N]$, s katerimi opišemo vse časovne položaje v signalu, ki predstavljajo možna mesta začetkov posameznih segmentov. Množica V vsebuje samo en izhodni simbol, ki določa mejo med dvema segmentoma, zaradi česar je v nadaljevanju izpuščena. Množica začetnih verjetnosti stanj sistema je predstavljena s:

$$\Pi = \begin{cases} \frac{1}{\eta},\ i \in [1, \eta], \\ 0,\ i \in [\eta, N], \end{cases} \tag{5.10}$$

kjer η predstavlja primerno izbrano vrednost za dolžino časa od začetka signala, kjer pričakujemo pričetek pesmi. V tem delu je lahko tišina ali morebiti tudi neuspel poskus začetka izvedbe.

Matrike verjetnosti prehodov med stanji A in izhodnih emisijskih vrednosti B natančneje predstavimo v nadaljevanju. Takšen model je predstavljen na sliki 5.10.

Cilj segmentacije je v signalu poiskati optimalno zaporedje stanj $S_{1:Q}$, glede na podane verjetnosti posameznega stanja $s_t = P(S_t)$ in verjetnost prehoda iz enega stanja v neko drugo stanje $a_{t-1,t} = P(S_t|S_{t-1})$. To dosežemo z maksimizacijo verjetnosti



Slika 5.10: Primer markovskega modela, ki ponazarja uporabljen verjetnostni model.

prehodov med stanji, definirano s spodnjo enačbo:

$$P(S_{1:Q}) = P(S_1) \prod_{t=2}^{Q} P(S_t|S_{t-1})P(S_t). \tag{5.11}$$

5.5.1 Verjetnosti stanj

Verjetnosti posameznih stanj $b_i = P(S_t = s_i)$ modeliramo z emisijskimi verjetnostmi modela v množici E. Verjetnost b_i je proporcionalna verjetnosti, da se meja med dvema segmentoma nahaja v času t_i. Sledimo razmisleku, da je verjetnost pojavitve meje med segmentoma na nekem mestu v signalu večja, v kolikor se pred tem mestom nahaja območje nizke amplitude signala. Pri petju se to največkrat odraža kot dihalna pavza, v instrumentalni glasbi pa kot zaključek neke glasbene fraze. Prav tako predpostavimo, da daljše kot je to območje nizke amplitude v signalu, tem večja je verjetnost pojavitve meje med segmentoma.

Iskanje območij nizke amplitude

V posnetkih komercialne glasbe je območja nizke amplitude precej preprosto zaznati, saj so zaradi kvalitete snemanja in naknadne obdelave posnetkov vsa takšna območja signala pod nekim amplitudnim pragom, ki ga lahko določimo tudi globalno. Takšne tehnike se ne moremo poslužiti pri ljudski glasbi, saj je kvaliteta posnetkov zelo slaba. Prav tako amaterski izvajalci pogosto pojejo s spremenljivo glasnostjo, občasno tudi

pod želenim globalnim pragom, ki ga določimo glede na amplitudo prisotnega šuma. V našem primeru to rešujemo z adaptivnim pragom.

Najprej izračunamo amplitudno ovojnico zvočnega signala A_e, ki je na primeru na sliki 5.11 označena s svetlo modro barvo. Ovojnico izračunamo tako, da najprej magnitudo signala filtriramo z nizkoprepustnim Butterworthovim filtrom 4. stopnje z mejno frekvenco 110 Hz, jo pretvorimo v decibele in prevzorčimo z vzorčevalno frekvenco 10 Hz.

Nadalje izračunamo adaptivno oceno povprečne amplitude skozi čas, s tem da amplitudno ovojnico filtriramo z medianinim filtrom 5. stopnje in dobimo (A_a). Ocena povprečne amplitude signala je na sliki 5.11 označena s črno barvo.

Na tem mestu se soočimo s problemom definiranja praga za definicijo območij nizke amplitude v signalu $(A_e < A_a + A_t)$. Ker uporaba fiksnega praga ne pride v poštev zaradi prisotnosti odzadnjega šuma v signalu, ki lahko dosega visoko stopnjo, definiramo t. i. adaptivni prag s spodnjo mejo. Spodnjo mejo postavimo na -50 dB, kar je vrednost, ki se večinoma obravnava za tišino v signalu. Kadar amplitudna ovojnica presega vrednost spodnje meje, je vrednost praga določena kot minimalna vrednost odmika od amplitudne ovojnice na intervalu $[-10\text{ dB}, -6\text{ dB}]$, ko območja nizke amplitude zajemajo vsaj 10 % dolžine celotnega signala, kar se izkaže za smiselno oceno glede na pregledane materiale v glasbeni zbirki. Ta območja se odražajo v prekinitvah in dihalnih pavzah v posnetku. S tem poskrbimo, da neničelna stanja v našem modelu niso prekomerno razpršena in posledično preprečimo prekomerno podsegmentacijo.

Emisijske verjetnosti modela

Emisijsko verjetnost posameznega stanja modela b_i določimo kot sorazmerno velikosti območja z nizko amplitudo v signalu neposredno pred časom t_i, pod pogojem, da takšno območje res obstaja in se konča prav v času t_i.

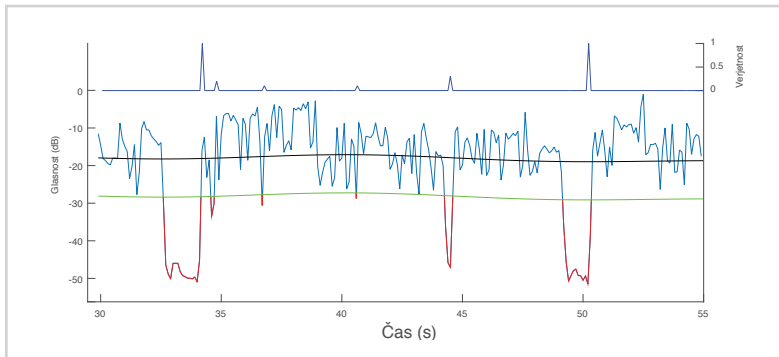
$$b_i = \begin{cases} 0,5;\ i = 1, \\ \tau;\ A_e > A_a + A_t \\ \propto \min(1, L_{l,i});\ A_e < A_a + A_t, \\ 1;\ i = N. \end{cases} \tag{5.12}$$

Emisijska verjetnost posameznega stanja modela b_i je sorazmerna dolžini območja z nizko amplitudo $L_{l,i}$ pred časom t_i.

Da preprečimo prevelik vpliv dolgih območij z nizko amplitudo signala, navzgor omejimo vrednost prispevka k emisijski verjetnosti stanja s prispevkom, ki ga določa območje dolžine ene sekunde. To pomeni, da imajo območja nizke amplitude, daljša ali enaka eno sekundo, na emisijsko verjetnost v nekem stanju enak vpliv. Emisijske verjetnosti stanja so na primeru prikazane na sliki 5.11 s temno modro barvo.

Emisijsko verjetnost prvega stanja, ki pred sabo ne mora imeti območij nizke amplitude, nastavimo na vrednost $0,5$, emisijsko verjetnost zadnjega stanja pa na vrednost 1, saj želimo, da se segmentacija konča v zadnjem stanju.

Celoten proces je prikazan tudi na sliki 5.11, kjer je amplitudna ovojnica signala A_e prikazana s svetlo modro barvo, območja ovojnice z nizko amplitudo so označena z rdečo barvo, povprečna amplituda A_a je prikazana s črno barvo, izbran prag $A_a + A_t$ je označen z zeleno barvo in emisijske verjetnosti stanj s temno modro barvo s skalo na desni strani.



Slika 5.11: Amplitudna ovojnica signala (svetlo modra, rdeče pod pragom), povprečna amplituda (črna), izbran prag (zelena) in verjetnosti stanj (temno modra).

5.5.2 *Izračun verjetnosti prehodov med stanji*

Verjetnost prehodov med stanjema i in j, $a_{ji} = P(S_t = s_i|S_{t-1} = s_j)$, določa, s kakšno verjetnostjo se bo naslednja meja med segmenti nahajala v času t_i, če je bila prejšnja v času t_j. Pri verjetnosti prehodov upoštevamo tri smiselne omejitve:

- segmenta z začetki v časih t_i in t_j morata biti med seboj podobna, kar ugotovimo iz povprečne podobnostne krivulje d_a;
- segmenta morata biti oddaljena približno za ocenjeno dolžino segmenta L, kot smo ga izračunali v predhodnem koraku. Dopuščena so odstopanja za $\sigma = \frac{L}{2}$;
- dovoljeni so samo prehodi naprej v času.

To lahko predstavimo tudi kot:

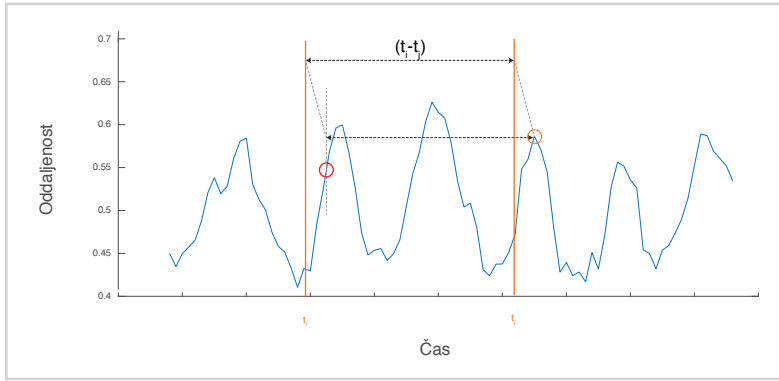
$$P(S_t = s_i|S_{t-1} = s_j) \propto \text{sim}(s_j, s_i)\mathcal{N}(L, \sigma),$$
$$P(S_t = s_i|S_{t-1} = s_j) = 0; t_i \leq t_j.$$

$\mathcal{N}$ predstavlja normalno porazdelitev. Funkcija podobnosti sim je izračunana iz obratne normalizirane krivulje oddaljenosti $|d_a^{'}|$, definirane kot:

$$|d_a^{'}| = 1 - \frac{d_a - \min(d_a)}{\max(d_a)}, \tag{5.13}$$

za katero velja, da zavzema vrednosti na intervalu $[0, 1]$, vrhovi funkcije pa predstavljajo ponovitve.

Ker takšna krivulja ni absolutno locirana v času, saj je nastala na podlagi naključno izbrane lokacije v signalu, pridobimo podobnost med segmentoma, ki se začneta v časih t_i in t_j, tako, da v krivulji poiščemo vrh, ki je najbližji času t_j (na sliki 5.12 oranžno obkrožena vrednost krivulje), in za podobnost vzamemo vrednost v časovnem odmiku $(t_i - t_j)$ od tega vrha (na sliki 5.12 rdeče obkrožena vrednost krivulje). Vrhovi predstavljajo ponovitve, in ker naj bi se ponovitve odražale v začetkih segmentov, s tem modeliramo podobnost segmenta, ki se začne v času t_j, glede na segment, ki se začne v času t_i. Celoten postopek je prikazan na sliki 5.12.



Slika 5.12: Prikaz, kako pridobimo podobnost med izbranima segmentoma.

Za iskanje optimalnega zaporedja stanj v modelu uporabimo Viterbijev algoritem. Pri iskanju optimalnega zaporedja stanj dopustimo, da je začetno stanje znotraj η sekund od začetka posnetka in model prisilimo, da se zaključi v končnem stanju. Ker se stanja neposredno preslikajo v čas, predstavlja končno zaporedje stanj ravno množico odkritih mej med segmenti.

6

Ovrednotenje razvite metode

Trying to understand the way nature works involves a most terrible test of human reasoning ability. It involves subtle trickery, beautiful tightropes of logic on which one has to walk in order not to make a mistake in predicting what will happen. The quantum mechanical and the relativity ideas are examples of this.

– Richard P. Feynman

V tem poglavju predstavimo ovrednotenje novorazvite metode za segmentacijo ljudskih pesmi. Poleg ovrednotenja natančnosti metode na zbirki ljudskih pesmi smo metodo ovrednotili tudi s stališča robustnosti delovanja v odvisnosti od degradacije posnetka.

6.1 Uspešnost segmentacije

Predstavljeno metodo smo ovrednotili na že predstavljeni zbirki ljudske glasbe. V tabeli 6.1 so prikazani rezultati ovrednotenja predlaganega pristopa v primerjavi z izbranimi aktualnimi metodami. V tabeli so povprečne vrednosti natančnosti, priklica in mere F1 za pesmi glede na posamezni sestav in celotno zbirko. Izračunana meja med segmenti šteje za pravilno (angl. true positive), če se nahaja znotraj ± 3-sekundnega okna okoli anotirane meje (enaka mera se uporablja tudi pri vrednotenju algoritmov MIREX).

Pri ovrednotenju predlagane metode smo uporabili naslednje vrednosti parametrov:

- povprečna dolžina kitice v pesmih $l = 20$ sekund;
- dovoljeno drsenje višine tonov $\zeta = 2$ poltona;
- minimalna vrednost stanja $\tau = 0,01$;
- širina normalne porazdelitve za kaznovanje pri odstopanju od pričakovane dolžine segmenta $\sigma = \frac{L}{2}$;
- množica dovoljenih začetnih stanj verjetnostnega modela $\eta = 6$ sekund.

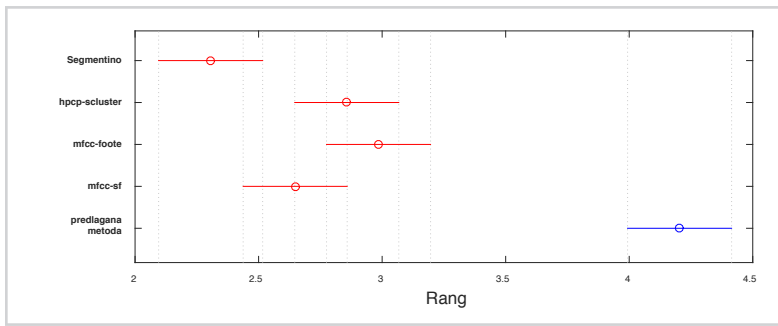
Vrednosti so bile izbrane glede na izkušnje in niso bile optimizirane specifično za zbirko pesmi. Z dodatnimi testi smo preverili, da predlagana metoda ni zelo občutljiva na spremembe teh vrednosti.

Predlagan pristop se na predstavljeni zbirki pesmi odreže signifikantno bolje za neinstrumentalne posnetke, rezultati na instrumentalnih posnetkih pa so primerljivi z najboljšo metodo - Segmentino. Ker smo se pri načrtovanju naše metode osredotočili na neinstrumentalne posnetke, je takšen rezultat tudi pričakovan. Rezultati našega pristopa so tudi najbolj uravnovešeni glede na natančnost in priklic, kar pomeni, da predlagan pristop pretirano ne podsegmentira ali nadsegmentira zvočnih posnetkov. Signifikantno izboljšanje pokaže tudi Friedmanov statistični test, katerega rezultati so

Tabela 6.1: Rezultati ovrednotenja predlagane metode v primerjavi z izbranimi aktualnimi metodami na zbirki ljudskih pesmi. Mere natančnost (P), priklic (R) in F1 so izračunane kot povprečna vrednost mer pesmi posameznega sestava.

Sestav		Segmentino	MSAF-MFCC-Foote	MSAF-HPCP-SCluster	MSAF-MFCC-SF	Müller	Predlagana metoda
Solo (OGL)	P	0,86	0,38	0,41	0,36		0,84
	R	0,17	0,76	0,51	0,42		0,85
	F1	0,25	0,51	0,46	0,39	0,87	0,85
Dvo- troglasje	P	0,82	0,48	0,44	0,52		0.84
	R	0,27	0,85	0,50	0,60		0.89
	F1	0,33	0,61	0,47	0,55		0,84
Instrumental	P	0,55	0,31	0,38	0,33		0,69
	R	0,82	0,97	0,87	0,80		0,63
	F1	0,62	0,47	0,53	0,47		0,60
Instr. - petje	P	0,55	0,37	0,35	0,42		0,74
	R	0,72	0,86	0,64	0,76		0,62
	F1	0,59	0,52	0,46	0,54		0,61
Solo	P	0,92	0,46	0,45	0,46		0,87
	R	0,26	0,78	0,64	0,49		0,87
	F1	0,36	0,57	0,53	0,48		0,86
Zbor	P	0,74	0,31	0,40	0,40		0,73
	R	0,36	0,76	0,61	0,64		0,90
	F1	0,41	0,44	0,48	0,50		0,78
Skupaj	P	0,74	0,39	0,41	0,41		0,78
	R	0,40	0,81	0,59	0,56		0,80
	F1	0,40	0,52	0,48	0,47		0,76

prikazani na sliki 6.1. V Friedmanov test nismo mogli vključiti metode *Müller*, saj zanjo nimamo rezultatov za posamezen primer.



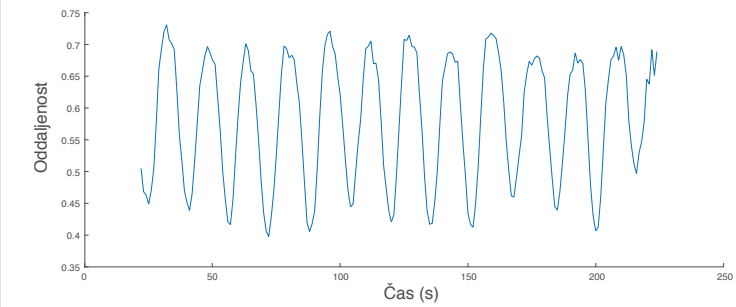
Slika 6.1: Rezultati Friedmanovega primerjalnega statističnega testa.

Rezultati predstavljenega pristopa so primerljivi tudi z najboljšimi metodami za segmentacijo ljudskih pesmi. Tako lahko primerjamo rezultate naše metode z rezultati

metode, predstavljene v [Müller et al., 2013], kjer svojo metodo testirajo na zbirki *Onder de groene linde* - Solo (OGL), ki so nam jo posredovali avtorji prispevka. Njihovi rezultati so minimalno boljši – mera F1 je 0,87 za zbirko Solo (OGL). Pri tem moramo izpostaviti, da avtorji uporabljajo F0 ojačane značilnice CENS [Müller, 2007], ki so posebej prilagojene za solo petje in posledično ne moremo oceniti, kako bi metoda delovala na zvočnih posnetkih drugih sestavov.

Analiza rezultatov posameznih pesmi je razkrila, da popolno natančnost in priklic (1,0) naša metoda doseže za 86 pesmi. Za šest pesmi naša metoda povsem odpove z mero F1 enako nič. Vrednost natančnosti je enaka nič v 10 pesmih, za katere velja, da naša metoda v polovici primerov pravilno ugotovi dolžino segmenta, zgreši pa absolutni začetek kitic, zaradi česar pride do t. i. faznega zamika in je posledično segmentacija v celoti napačna.

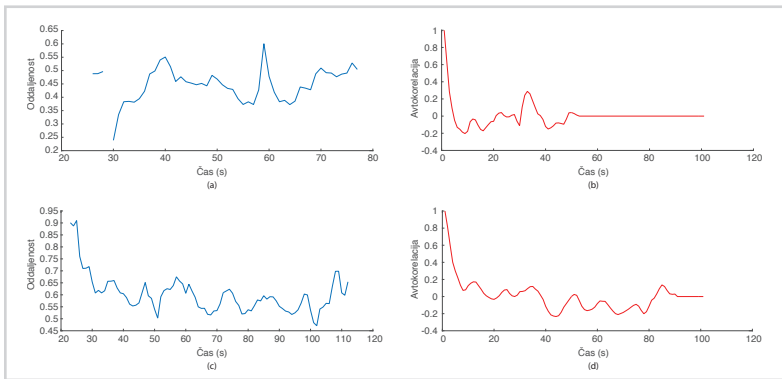
Kadar je povprečna krivulja oddaljenosti pravilno ocenjena, je segmentacija večinoma pravilna. Primer krivulje, kjer so ponovitve razločno vidne kot lokalni minimumi na krivulji oddaljenosti, je prikazan na sliki 6.2.



Slika 6.2: Povprečna krivulja oddaljenosti, kjer so ponovitve izrazito vidne kot lokalni minimumi.

Pravilna segmentacija je v naši metodi odvisna tudi od pravilno ocenjene dolžine segmenta. V nekaterih primerih je naš pristop izbral napačno dolžino segmenta in posledično nepravilno izvedel segmentacijo. Takšnih primerov je 27, a tudi za nekatere med njimi pristop najde vsaj nekaj pravilnih mej med posameznimi segmenti. Kadar posamezna kitica sestoji iz dveh podobnih delov, lahko ocenjena dolžina segmenta postane polovica dejanske dolžine segmenta glede na avtokorelacijsko krivuljo. Kot rezultat

dobimo nadsegmentirano pesem za faktor 2. Nadalje velja, da kadar so podobnosti znotraj posameznega segmenta visoke, kar je pogosto v instrumentalni glasbi, krivulja oddaljenosti ni izrazito periodična, kar se prav tako odraža v napačno ocenjeni dolžini posameznega segmenta in posledično v napačni segmentaciji. Dva takšna primera sta prikazana na sliki 6.3. Povprečna krivulja oddaljenosti nima izrazitih lokalnih minimumov (slika 6.3 (a)) in posledično z avtokorelacijo ne ocenimo pravilno predvidene dolžine segmenta (slika 6.3 (b)). V primeru, prikazanem na sliki 6.3 (c), je v krivulji oddaljenosti preveč izrazitih lokalnih minimumov. Posledično z avtokorelacijo napačno ocenimo dolžino segmenta, ki je prekratek za faktor 2, in posledica je dvakratna nadsegmentacija pesmi (slika 6.3 (d)).



Slika 6.3: Slika prikazuje dva primera, kjer predstavljena metoda odpove.

6.2 Ovrednotenje robustnosti metode

Metodo smo razvili z namenom, da bo delovala na posnetkih ljudske glasbe, za katere velja, da so lahko slabe kvalitete tako zaradi neprofesionalne snemalne opreme in slabih snemalnih pogojev kot tudi amaterskih izvajalcev. Že testna glasbena množica, na kateri smo metodo ovrednotili, vsebuje posnetke z omenjenimi lastnostmi, a smo se kljub temu odločili za dodatno sistematično testiranje performans pri različnih degradacijah.

Za ovrednotenje smo uporabili ogrodje za degradacijo zvočnih posnetkov (angl. audio degradation toolbox - ADT) [Mauch and Ewert, 2013], ki omogoča različne načine in stopnje degradacije zvočnih posnetkov. Degradacije, za katere smo izvedli ovrednote-

nje, so naslednje:

- *dodajanje šuma* - v posnetek doda naključno generiran šum različnih *barv* (tipov): rdeči, roza, beli, modri in vijolični. Pri tem lahko nadzorujemo razmerje signal-šum (angl. signal-to-noise ratio).
- *dodajanje zvoka* - posnetku doda zvočni posnetek v izbranem razmerju signal-šum.
- *prekrivanje* (angl. aliasing) - popačenje signala zaradi prevzorčenja na nižjo frekvenco brez nizkoprepustnega filtra.
- *rezanje* (angl. clipping) - efekt, do katerega pride, ko zvočni vir preseže maksimalno glasnost in se vrhovi signala porežejo.
- *kompresija dinamičnega razpona* (angl. dynamic range compression) - signal nad določeno mejo glasnosti stiša in s tem zmanjša razlike med tihimi in glasnimi deli posnetka.
- *visoko-/nizkoprepustni filter* - iz signala izloči frekvence pod/nad določenim pragom.
- *stiskanje mp3* - na signal aplicira efekt stiskanja mp3 za določene nastavitve bitne širine.
- *harmonično popačenje* - signal popači z dodajanjem harmoničnih komponent. Transformacijo lahko apliciramo večkrat zapored za dosego močnejše degradacije.

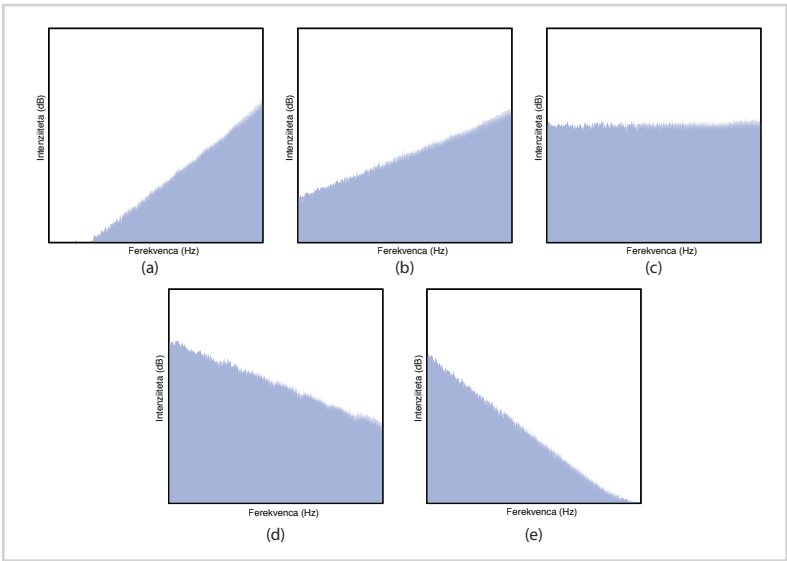
6.2.1 Dodajanje šuma

Pri degradaciji z dodajanjem šuma smo v signal dodajali različne vrste šuma (beli, roza, rdeči, modri in vijolični šum). Vsaka izmed vrst šuma ima svoje lastnosti in signal degradira na drugačen način:

- *beli šum:* predstavlja naključni signal s konstantno gostoto močnostnega spektra, katerega vzorci se obravnavajo kot zaporedje nepovezanih naključnih vrednosti s povprečjem nič in končno varianco.

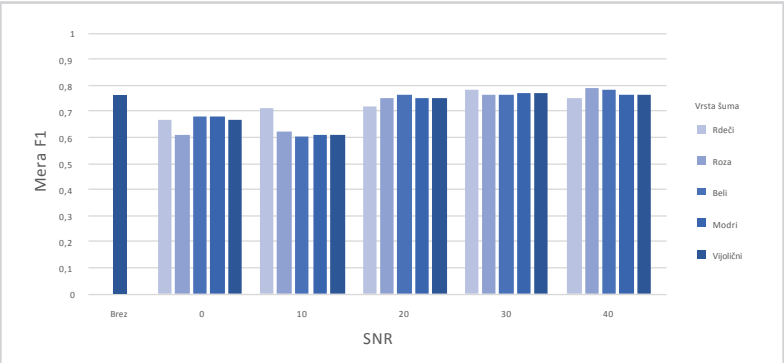
- *roza šum:* predstavlja signal, katerega frekvenčni spekter je takšen, da je gostota močnostnega spektra obratno sorazmerna frekvenci signala. V roza šumu ima vsaka oktava (razpolovitev/podvojitev frekvence) enak del moči signala.
- *rdeči šum (znan tudi kot Brownov šum):* predstavlja signal, kot ga proizvede Brownovo gibanje. Imenujemo ga tudi šum naključnega sprehoda. Ime rdeči šum izvira iz analogije svetlobi. Če je beli šum enakomerno porazdeljen po celotnem spektru, za rdeči šum velja, da ima večjo moč pri večjih valovnih dolžinah.
- *modri šum:* predstavlja naključni signal, ki ima ravno nasprotne lastnosti kot roza šum. Ta šum ima večjo moč pri krajših valovnih dolžinah.
- *vijolični šum:* predstavlja naključni signal, ki ima ravno nasprotne lastnosti kot Brownov šum. Večjo moč ima pri krajših valovnih dolžinah. Pri najkrajših je lahko njegova moč tudi nič.

Vsi šumi so predstavljeni na sliki 6.4.



Slika 6.4: Prikaz različnih vrst šuma: vijolični (a), modri (b), beli (c), roza (d) in rdeči (e).

Koliko šuma je v signalu, definiramo z razmerjem signal-šum. V primeru dodajanja šuma smo v signal dodajali šum v razmerjih 40, 30, 20, 10 in 0. Pri tem razmerje 40 predstavlja najnižjo stopnjo šuma, razmerje 0 pa najvišjo stopnjo šuma v signalu. Pri razmerju nič je moč originalnega signala ravno enaka moči dodanega šuma. Kot se izkaže, dodajanje šuma v signal ne vpliva močno na sam rezultat predstavljene metode, kar je razvidno iz rezultatov, predstavljenih na grafu na sliki 6.5.



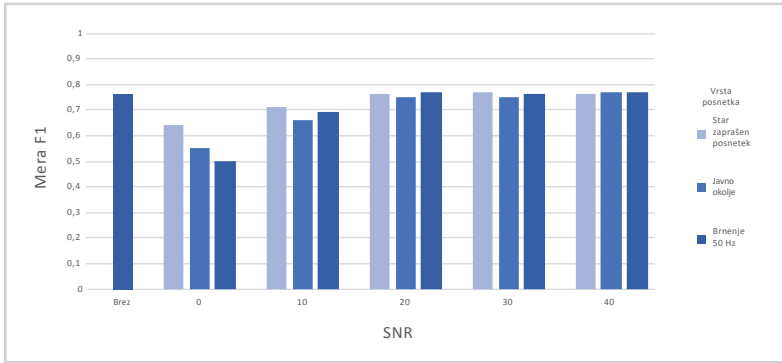
Slika 6.5: Odpornost metode na dodajanje šuma v signal.

Po pričakovanjih se rezultati metode ob dodajanju šuma nekoliko slabšajo. Kljub temu tudi v najslabšem primeru, ko v signal dodamo šum z razmerjem signal-šum nič, metoda še vedno ne odpove, rezultati so, glede na mero F1, slabši za 0,16, kar predstavlja 21 % poslabšanje. Kot se izkaže, vrne metoda v določenih primerih degradacije z dodajanjem šuma enake ali boljše rezultate kot osnovna metoda.

6.2.2 *Dodajanje zvoka*

Degradacija posnetkov z dodajanjem drugih zvočnih posnetkov predstavlja možnost postavitve obstoječega zvočnega vira v neko drugo okolje tako, da dodamo lastnosti zvočnega vira okolja ali zvočne scene. Izbrali smo tri vrste dodanega zvoka: *star zaprašen posnetek*, *javno okolje* in *brnenje pri 50 Hz*. Podobno kot pri dodajanju šuma tudi tukaj degradacijski posnetek dodajamo glede na razmerje signal-šum. Za razmerja vzamemo enake vrednosti kot pri degradaciji s šumom: 40, 30, 20, 10 in 0.

Dodajanje zvokov ima večji vpliv na delovanje metode kot dodajanje šuma, saj pri niž-



Slika 6.6: Odpornost metode na dodajanje posnetka v signal.

jem razmerju (razmerje signal-šum je nič) uspešnost metode pade tudi do 0,27, kar predstavlja 35,5 % poslabšanje, kar je razvidno z grafa na sliki 6.6. Kljub temu pa pri visokem razmerju (razmerje signal-šum je 40 do 20) uspešnost metode praktično ne pade. Ugibamo lahko, da je večji padec pri nižjem razmerju posledica dejstva, da v signal dodajamo isti zvok, kar pomeni, da dodamo samopodobnost v signal, kar posledično negativno vpilva na izračun podobnosti. Pri šumu zaradi naključnosti dodanega signala do teh problemov ne prihaja.

6.2.3 Prekrivanje

Pri prekrivanju zaradi vzorčenja signala z nižjo vzorčevalno frekvenco signal zajame manj podrobnosti, predvsem v delu spektra z višjimi frekvencami. Degradacija deluje tako, da se originalni posnetek prevzorči brez nizkoprepustnega filtra, kar povzroči, da se v signalu pojavijo neobstoječe frekvence. Za naše teste smo izbrali frekvence vzorčenja 11.050 Hz, 8.000 Hz in 4.000 Hz. Kljub temu, da je degradacija posnetka precej visoka, pa predlagana metoda ne vrača nič slabših rezultatov. Še več, pri vzorčenju s frekvenco 4.000 Hz se metoda odreže celo 4 % bolje, kar se odraža v povečanju vrednosti mere F1 za 0.02. Možna razlaga za boljše rezultate pri nizkih vzorčevalnih frekvencah je, da s tem iz signala izločimo visokofrekvenčni šum. Rezultati degradacije nakazujejo, da prisotnost visokih frekvenc poslabša delovanje metode, kar potrdijo tudi rezultati degradacije z nizkoprepustnim filtriranjem, predstavljeni v nadaljevanju.

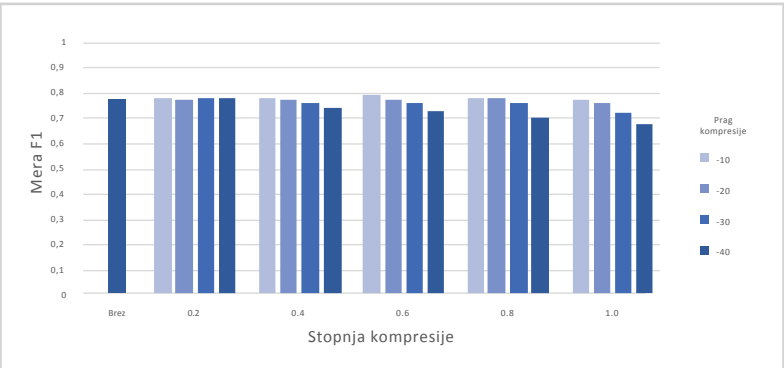
6.2.4 *Rezanje*

Pri degradaciji z efektom rezanja signalu odrežemo najglasnejši del. Rezultat je enak, kot bi do preseganja glasnosti prišlo že med samim snemanjem. Za testiranje smo izbrali, da bomo odrezali najglasnejši del signala, ki zajema 1 %, 2 %, 5 % in 10 % celotnega signala. Degradacija z efektom rezanja ne vpliva bistveno (manj kot 2 % spremembe) na samo delovanje metode, v nekaj primerih so rezultati celo boljši od osnovne metode.

S to degradacijo dosežemo v posnetku bolj enakomerno glasnost, šum, ki ga povzroči rezanje, pa na metodo ne vpliva bistveno, saj kromatične značilnice, uporabljene v metodi, izračunamo z dovolj dolgimi okni.

6.2.5 *Kompresija dinamičnega razpona*

Degradacija z uporabo kompresije dinamičnega razpona zmanjša dinamični razpon v signalu. To velja seveda tudi za območja signala s tišino, kar bi lahko vplivalo na samo delovanje metode. Kot se izkaže, sama degradacija nima večjega vpliva na delovanje metode. Parametra dinamične kompresije sta *prag kompresije* in *stopnja kompresije*. Za prag kompresije smo izbrali vrednosti -10, -20, -30 in -40 dB, za stopnjo kompresije pa vrednosti 0,2, 0,4, 0,6, 0,8 in 1,0.



Slika 6.7: Odpornost metode na degradacijo dinamične kompresije.

Kot je razvidno iz rezultatov, predstavljenih na grafu na sliki 6.7, predstavljena meto-

da ni zelo občutljiva na degradacijo dinamične kompresije. V najslabšem primeru se vrednost mere F1 poslabša za 0,10, kar predstavlja 13 % poslabšanje, v veliko primerih se sploh ne poslabša ali pa se poslabša zgolj za nekaj odstotkov.

6.2.6 *Visokoprepustno filtriranje*

Degradacija posnetka z visokoprepustnim filtrom v posnetku zmanjša magnitudo signala pri nizkih frekvencah in s tem seveda povzroči izbris določenih informacij. Pri degradaciji smo za mejno frekvenco visokoprepustnega filtra uporabili vrednosti 55, 110, 220 in 440 Hz. Na ta način odrežemo pomembne dele govorjenega oz. petega signala. Posledično v signalu ostanejo višje frekvence.

Degradacija z visokoprepustnim filtrom na delovanje predstavljene metode nima vpliva, saj so spremembe rezultatov manjše od 2 %. V določenih primerih so rezultati celo boljši kot brez degradacije.

6.2.7 *Nizkoprepustno filtriranje*

Nizkoprepustni filter v signalu zniža magnitudo visokih frekvenc, kar je ravno nasprotno od degradacije, predstavljene v predhodnem podpoglavju. Pri degradaciji z nizkoprepustnim filtrom smo za mejne frekvence izbrali 8.000, 4.000 in 2.000 Hz. S tem iz signala odstranimo predvsem višje harmonike osnovnih tonov.

Izkaže se, da ob takšni degradaciji posnetka predstavljena metoda vrača boljše rezultate (a ne za več kot 2 %), kar nakazuje na to, da je v posnetkih prisotnega nekaj visokofrekvenčnega šuma, sama informacija, potrebna za dobro delovanje metode, pa leži v nižjih frekvencah od izbrisanih. Rezultati so konsistentni z rezultati degradacije s prekrivanjem pri vzorčenju s 4.000 Hz.

6.2.8 *Harmonična popačenost*

Degradacija harmonične popačenosti izvede preprosto transformacijo $x \leftarrow \sin(\pi x)$ nad vsakim vzorcem signala, definiranim na intervalu [-1,1]. Z večkratnim ponavljanjem transformacije simuliramo efekt zasičenja (angl. saturation), saj se vrednosti vzorca postopoma pomikajo proti mejnim vrednostim -1 in 1. Pri našem testiranju

robustnosti smo se odločili, da bomo transformacijo nad posnetkom izvedli enkrat do petkrat. Izkaže se, da harmonično popačenje praktično ne vpliva na samo delovanje metode, saj so rezultati po popačenju praktično nespremenjeni (manj kot 1 % razlike).

6.2.9 Stiskanje mp3

Zadnja degradacija, ki smo jo izvedli, je popačenje, skladno s stiskanjem mp3. Pri tem stiskanju prihaja do različnih izgub informacij glede na delovanje samega kompresijskega algoritma. Kompresija se določa z želenim izhodnim podatkovnim tokom. V našem primeru smo izbrali vrednosti 128, 96 in 64 kb/s. Izkaže se, da sama kompresija mp3 na delovanje ne vpliva, saj so odstopanja v rezultatih ± 1 %.

6.2.10 Sklep o robustnosti

Na podlagi rezultatov po izvedenih degradacijah lahko ugotovimo, da razvita metoda na degradacije v signalu ni zelo občutljiva. Predstavljena metoda v nobenem primeru povsem ne odpove, saj se v najslabšem primeru rezultati mere F1 poslabšajo za 0,27 oz. 35,5 %. Zavedati se je potrebno, da so sami posnetki zelo slabi in so že bili podvrženi različnim degradacijam signala.

Že pri načrtovanju metode smo težili k njeni robustnosti z izbiro dolgih oken pri izračunu kromatičnih značilnic, z adaptivno detekcijo tišine, odstranjevanjem izstopajočih značilnic ipd. Želeno robustnost nakazujejo tudi rezultati predstavljenega ovrednotenja, saj je razvita metoda izredno robustna in neobčutljiva na najrazličnejše degradacije.

Zaključki

7

The first footfalls on Mars will mark a historic milestone, an enterprise that requires human tenacity matched with technology to anchor ourselves on another world.

– Buzz Aldrin

V prvem delu disertacije smo podrobno predstavili problematiko segmentacije glasbe, še posebej ljudske glasbe, kjer splošni segmentacijski pristopi ne vračajo zadovoljivih rezultatov. Najprej smo na zbirki ljudske glasbe iz arhiva Etnomuza ovrednotili trenutne segmentacijske pristope in izpostavili njihove slabosti. Glavne slabosti trenutno aktualnih segmentacijskih pristopov so v tem, da ne naslavljajo specifik ljudske glasbe, kot so visoka stopnja šuma, slabi snemalni pogoji, amaterski pevci ipd.

Na podlagi teh ugotovitev smo razvili lastno metodo, ki upošteva tudi lastnosti ljudske glasbe in se zato v segmentaciji ljudske glasbe odreže veliko bolje kot aktualni pristopi. Pri tem se nismo omejili na specifične sestave v ljudski glasbi (npr. solo petje), kot je predstavljeno v [Müller et al., 2013], ampak smo naslovili ljudsko glasbo različnih sestavov: solo petje, dvo- in triglasno petje, instrumental in mešani sestavi.

Razvita metoda na predstavljeni glasbeni zbirki dosega boljše rezultate kot trenutno aktualne metode. Na celotni zbirki najboljšo izmed ovrednotenih metod, ki doseže vrednost mere F1 0,52, naša metoda prekaša za več kot 46 % in doseže vrednost mere F1 0,76. Del zbirke, kjer delovanja aktualnih segmentacijskih metod nismo presegli, predstavlja instrumentalna glasba. Na instrumentalni glasbi nas prekaša metoda Segmentino, ki doseže za 3 % boljše rezultate, ki pa glede na Friedmanov statistični test niso signifikantno boljši. Podobno velja tudi za del zbirke solo petja iz glasbenega arhiva OGL, kjer aktualna metoda *Müller* doseže 2 % boljše rezultate od naše metode, a pri njej statistične signifikance nismo mogli preveriti, ker nimamo rezultatov za posamezno pesem.

Nenazadnje smo razvito metodo ovrednotili tudi z vidika robustnosti, kjer smo ugotovili, da je odporna tudi na velike degradacije zvočnega signala. Pri ovrednotenju smo uporabili 8 vrst degradacij, med katerimi je na delovanje metode najbolj vplivalo dodajanje zvočnega posnetka v signal. Vseeno pa se tudi pri najvišji stopnji te degradacije, kjer smo v obstoječe posnetke dodali zvok s frekvenco 50 Hz v razmerju signal-šum nič, rezultat metode poslabša zgolj za 21 % in doseže vrednost mere F1 0,6.

S predstavljeno metodo smo zadostili zastavljenemu cilju, predstavljenem v dispoziciji doktorske disertacije, kjer smo si zadali dosego izvirnega znanstvenega prispevka:

Robusten algoritem za segmentacijo ljudske glasbe, ki naslavlja probleme ljudskih pesmi.

7.1 Nadaljnje delo

Kljub dobremu delovanju metode pa predlagamo nekaj možnih izboljšav:

- V določenih primerih metoda povsem odpove, ker ne odkrije pravilnega začetka pesmi oz. ga zamakne, kar povzroči fazni zamik vseh mej med posameznimi segmenti. Izboljšava ocene verjetnosti začetkov segmentov bi lahko te zamike odpravila.
- Metoda v nekaj primerih za večkratnik napačno izračuna dolžino reprezentativnega segmenta, kar privede do pod- oz. nadsegmentacije. Problem bi lahko naslovili tako, da bi dodatno preverjali podobnosti znotraj reprezentativnega segmenta in na podlagi tega poskušali oceniti, ali je prišlo do katerega izmed omenjenih pojavov. V takšnem primeru bi lahko dolžino najbolj reprezentativnega segmenta ponovno ocenili.
- Predstavljena metoda ni prilagojena za boljše delovanje na kateri izmed specifičnih zvrsti glasbe, zastavljena je splošno. Za boljše rezultate na posameznem tipu gradiva bi lahko v metodo vpeljali še prepoznavo tipa gradiva (npr. [Marolt, 2009]), v nadaljnjih korakih metode pa bi uporabili nabor parametrov, ki so primernejši za posamezen tip gradiva.

Del II

Transkripcija zvočnih posnetkov ljudske glasbe

V drugem delu disertacije predstavljamo novo metodo za transkripcijo polifonične glasbe, ki smo jo razvili z namenom izboljšanja transkripcije polifoničnih posnetkov ljudske glasbe v primerjavi z rezultati trenutno aktualnih transkripcijskih metod.

Pod transkripcijo pojmujemo postopek pretvorbe zvočnega zapisa glasbe v simboličen zapis, kjer iz posnetka izločimo višino, začetek in trajanje tonov. Tovrsten simboličen zapis podaja torej višjenivojski opis izvedbe glasbenega dela, ki je uporaben pri analizi, organizaciji in iskanju ter poustvarjanju dela. Etnomuzikologi in muzikologi pri analizi zvočnega gradiva izdelajo tudi natančno transkripcijo, pri čemer navadno transkribirajo najbolj tipično interpretacijo ponavljajočega dela (kitice) v posnetku. Transkripcijo izvajajo ročno, kar je časovno zelo zamuden postopek. Avtomatizacija tega opravila je tako eden glavnih izzivov na področju pridobivanja informacij iz glasbe.

Razvita metoda temelji na podobnih predpostavkah kot predhodno predstavljen postopek za segmentacijo posnetkov ljudskih pesmi. Glavna od teh je, da je posnetek sestavljen iz ponavljajočih delov, kar izkoristimo za razvoj algoritma za izboljšanje ocene tonov, ki jih izračunamo z uporabo obstoječih transkripcijskih metod. Metodo smo ovrednotili na zbirki večglasnih ljudskih pesmi in rezultate primerjali z rezultati aktualnih transkripcijskih metod.

V poglavjih, ki sledijo, najprej predstavimo zbirko zvočnih posnetkov, ki smo jih uporabili za izvedbo ovrednotenja, nadalje predstavimo ovrednotenje izbranih aktualnih transkripcijskih metod na predstavljeni zbirki ter izpostavimo prednosti in pomanjkljivosti posameznega pristopa. Sledi predstavitev razvite metode in primerjava rezultatov z rezultati aktualnih metod. Za konec drugega dela podamo še sklepne ugotovitve, možne izboljšave in prilagoditve metode.

Zbirka večglasnih ljudskih pesmi

Vse pesmi za ovrednotenje postopkov smo izbrali iz zbirke Etnomuza. Ročne transkripcije vzorčnih kitic so izdelali muzikologi na GNI ZRC SAZU, poravnavo in transkripcijo celotnih zvočnih posnetkov pa je na podlagi teh vzorčnih kitic izvedel Paščinski v okviru diplomskega dela [Paščinski, 2015]. Izbrali smo 37 večglasnih pesmi v skupnem trajanju 107,6 minut. Povprečna pesem je dolga 174 sekund in ima povprečno 8,7 kitice, povprečna kitica pa je dolga 21 sekund.

Ovrednotenje aktualnih transkripcijskih metod

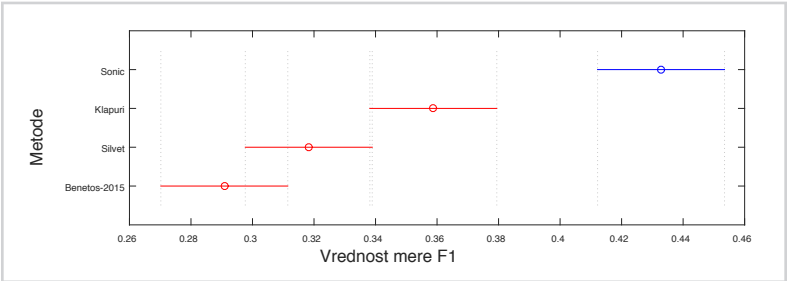
8

The whole of science is nothing more than a refinement of everyday thinking. It is for this reason that the critical thinking of the physicist cannot possibly be restricted to the examination of the concepts of his own special field.

– Albert Einstein

Za analizo uspešnosti transkripcijskih metod na večglasni ljudski glasbi smo zbrali nekaj javno dostopnih aktualnih transkripcijskih metod za polifonično glasbo in jih ovrednotili. Izbrali in ovrednotili smo naslednje algoritme, ki lahko podajo oceno višin osnovnih tonov v signalu: Sonic [Marolt, 2004], Klapuri [Klapuri, 2003], Silvet [Benetos and Weyde, 2013] in Benetos-2015 [Benetos and Weyde, 2015].

Vsako izmed metod smo ovrednotili na celotni zbirki in izračunali povprečno vrednost mere F1. Pri tem smo preizkusili različne nabore parametrov metod in izbrali najboljše. Medsebojna primerjava uspešnosti metod je prikazana na sliki 8.1. Pri evalvaciji primerjamo ujemanje napovedanih višin tonov glede na ročno transkripcijo za posamezno časovno okno.



Slika 8.1: Primerjava transkripcijskih metod glede mero F1 z označenimi intervali zaupanja pri 95 % stopnji.

Evalvacija se je izvajala na 40 ms oknih, kjer smo za pravilno transkripcijo šteli tisto transkripcijo posameznih glasov, pri kateri so bili glasovi največ četrtino tona oddaljeni od ročne anotacije. Rezultati ovrednotenja omenjenih metod so z merami natančnost (P), priklic (R) in F1 predstavljeni tudi v tabeli 8.1, kjer so z rdečo obarvani najboljši rezultati.

Tabela 8.1: Rezultati ovrednotenja izbranih aktualnih transkripcijskih metod na predstavljeni zbirki ljudskih pesmi. Mere natančnost (P), priklic (R) in F1 so izračunane kot povprečna vrednost mer za posamezno pesem.

Metoda	P	R	F1
Sonic	0,39	0,50	0,43
Klapuri	0,26	0,59	0,36
Silvet	0,25	0,46	0,32
Benetos-2015	0,21	0,47	0,29

V nadaljevanju predstavljamo analizo rezultatov izbranih metod.

8.1 Sonic

Metodo Sonic je Marolt predstavil v delu [Marolt, 2004] in je bila razvita za transkripcijo polifonične klavirske glasbe. Kljub temu, da je metoda prilagojena za transkripcijo točno določenega instrumenta, se v primerjavi z ostalimi metodami pri transkripciji večglasne vokalne glasbe dobro izkaže. Sama metoda je sestavljena iz treh delov. V prvem delu je vhodni signal analiziran s pomočjo avditornega modela, ki simulira delovanje človeškega ušesa. Sestavljata ga zbirke filtrov za razcep signala v več frekvenčnih kanalov in model, ki simulira aktivacije nevronov v slušnem živcu. Drugi del metode uporablja mreže adaptivnih oscilatorjev za odkrivanje in sledenje harmoničnim kompleksom v signalu. V tretjem delu so izhodi mrež adaptivnih oscilatorjev uporabljeni kot vhod v nevronske mreže za prepoznavanje not v polifoničnem signalu. Zbirka 76 nevronskih mrež je naučena na zbirki klavirske glasbe.

Metoda se med vsemi testiranimi metodami obnese najbolje glede na vrednost mere natančnosti in F1 ter ima med vsemi testiranimi metodami drugo najvišjo vrednost mere priklica. Razlog za to, da rezultati niso še boljši, je po vsej verjetnosti ravno v dejstvu, da so nevronske mreže, uporabljene v pristopu, učene, na zbirki klavirske glasbe, kjer višji harmoniki niso močno izraženi, kar pa pri petju ne velja in vodi do odkrivanja velike količine visokih tonov. Metoda se obnese bolje od ostalih najverjetneje tudi zaradi tega, ker v zadnjem koraku tvori posamezne note in ne le ocen osnovnih tonov kot ostale metode. Zaradi tega izhod metode vsebuje manj šuma v obliki osamljenih ocen osnovnih tonov z nizko stopnjo podobnosti.

8.2 Klapuri

Klapurijev pristop [Klapuri, 2003] za polifonično transkripcijo uporablja metodo, ki temelji na redukciji spektra. Pristop ciklično ponavlja naslednje korake: (1) oceni najmočnejšo višino tona v posameznem okvirju signala, (2) na podlagi harmoničnosti zajame del signala z zaznano višino tona, (3) iz signala odstrani del, ki predstavlja zaznano višino tona, (4) oceni število preostalih zvočnih virov in se vrne na prvi korak.

Metoda se po vrednosti mere F1 uvrsti na drugo mesto. Med vsemi testiranimi metodami ima najboljšo vrednost priklica, to je 0,59, pri natančnosti pa se uvršča na drugo mesto. Razlog za takšno delovanje je podoben kot pri Sonicu. Metoda predpostavlja, da amplituda harmonične serije eksponentno pada pri višjih harmonikih, kar pri petju ne drži. Tako pri odstranjevanju tonov odstrani premajhen delež višjih harmonikov, kar vodi do odkrivanja novih neobstoječih tonov.

8.3 Silvet

Metoda Silvet [Benetos and Dixon, 2013] za avtomatsko transkripcijo večglasja modelira časovni razvoj posameznega tona v signalu. Predstavljeni model razširja metodo SI-PLCA s podporo za spektralne predloge, ki ustrezajo stanjem zvoka: pričetku (angl. attack), vzdržanosti (angl. sustain) in pojenjanju (angl. decay). Vpliv teh predlog nadzorujejo časovne omejitve, modelirane s HMM. Dodatno podpira model uporabo več predlog na višino tona in zvočni vir. HMM se uporablja tudi v zadnjem koraku metode za sledenje notam. Razviti model je treniran na predlogah različnih orkestralnih instrumentov. Metoda je bila najboljša na evalvaciji MIREX 2013 za nalogo *Ocena več višin osnovnih tonov in njihovo sledenje* (angl. multiple fundamental frequency estimation & tracking results).

Na predstavljeni zbirki zvočnih posnetkov ima predzadnjo vrednost natančnosti in mere F1 ter najslabšo vrednost priklica. Kot pri ostalih metodah lahko tako delovanje med drugim pripišemo tudi dejstvu, da je bila metoda prilagojena posnetkom orkestralnih instrumentov in s tem ne povzema lastnosti vokalov. Prav tako velja, da so spektralne značilnosti stanj zvoka (pričetek, vzdržanost in pojenjanje) za vokal precej drugačni kot za instrumente in posledično uporabljene spektralne predloge niso ustrezne.

8.4 Benetos-2015

Poleg Benetosove metode Silvet smo testirali tudi njegovo najnovejšo metodo Benetos-2015, predstavljeno v delu [Benetos and Weyde, 2015]. Pristop je bil razvit z idejo splošne transkripcije polifoničnih zvočnih posnetkov. Podobno kot predhodno predstavljena metoda tudi nova metoda temelji na uporabi PLCA in podpira uporabo spektralnih predlog za stanja zvoka. Za vhod je uporabljena časovno-frekvenčna predsta-

vitev s spremenljivo Q transformacijo (angl. variable Q transform). Za definiranje zvočnih stanj sta predstavljena dva pristopa: (1) brez časovnih omejitev in (2) omejitve, določene z modelom HMM. Metoda je bila najboljša na evalvaciji MIREX 2015 za nalogo *Ocena več višin osnovnih tonov in njihovo sledenje*.

Na naši glasbeni zbirki metoda ne deluje najbolje, saj je tako v meri F1 kot v natančnosti zadnja med testiranimi metodami, v meri priklica pa predzadnja. Dosega tudi slabše rezultate od avtorjeve prejšnje in enostavnejše metode Silvet. Razlog za slabe rezultate najverjetneje tiči v dejstvu, da je metoda izredno prilagojena za transkripcijo instrumentalnih posnetkov, predvsem z izborom spektralnih predlog za stanje zvoka.

8.5 Diskusija

Uspešnost predstavljenih metod kljub dejstvu, da so med najboljšimi metodami za transkripcijo polifoničnih zvočnih posnetkov, ni visoka. Najboljša metoda doseže vrednost mere F1 komaj 0,43, kar je daleč od želenega. Veliko pojasni dejstvo, da gre za posnetke ljudske glasbe, ki predstavljajo precejšen izziv za metode MIR, kot smo pojasnili že v uvodu te disertacije. Poleg tega se moramo zavedati, da tudi ročna transkripcija, uporabljena pri evalvaciji, ni popolna, saj so nihanja v višini tonov pri vokalnih izvedbah velika.

Med predstavljenimi metodami nobena ne izkorišča dejstva, da se v pesmih (in ne le pri ljudskih) segmenti pesmi ponavljajo. To dejstvo in želja po povečani robustnosti transkripcij ter upoštevanju specifik ljudske pesmi so nas vodile v razvoj lastne transkripcijske metode. Predhodno predstavljena metoda za segmentacijo nam bo pri tem v pomoč, saj je ravno izkoriščanje ponavljanj ena glavnih značilnosti našega pristopa.

Metoda za transkripcijo ljudskih pesmi

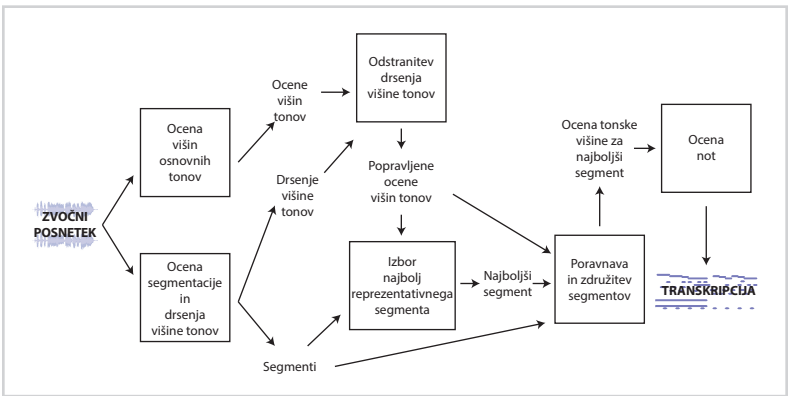
9

Don't keep forever on the public road,going only where others have gone, and following one after the other like a flock of sheep. Leave the beaten track occasionally and dive into the woods. Every time you do so you will be certain to find something that you have never seen before. Of course it will be a little thing, but do not ignore it. Follow it up, explore all around it; one discovery will lead to another, and before you know it you will have something worth thinking about to occupy your mind. All really big discoveries are the results of thought.

– Alexander Graham Bell

Transkripcijski pristop, razvit v okviru disertacije, vzame za vhod zvočni posnetek ljudske pesmi, ga procesira v več fazah in kot končni rezultat vrne seznam not z višinami ter začetnimi in končnimi časi. Posamezni koraki metode so prikazani na sliki 9.1. Metoda temelji na predpostavki, da je posamezna pesem sestavljena iz ponavljajočih segmentov. Ti se lahko med seboj razlikujejo v tempu, izvedbi zaradi netočnega izvajanja ali drsenja višine tonov, kot tudi zaradi spremenljivih snemalnih pogojev, kar se odraža v odzadnjem šumu ali prekinitvah. Ponavljajoča struktura je značilna za veliko večino ljudskih pesmi, kar smo izpostavili že v prvem delu disertacije. Pesmi so lahko večglasne (polifonične), metoda pa je še posebej primerna za transkripcijo vokalne glasbe.

Poudariti je potrebno, da metoda vrne najbolj reprezentativno transkripcijo pesmi, kot to počno tudi muzikologi, ki ne transkribirajo vseh ponovitev segmentov, ampak jih zanima najbolj tipično zapet segment pesmi in njegova transkripcija.



Slika 9.1: Metoda in njeni posamezni koraki.

Metoda združuje tri tipe časovno spremenljivih informacij za robusten izračun transkripcije:

1. višine osnovnih tonov, pridobljene z metodami, kot so Klapuri [Klapuri, 2003] ali Benetos [Benetos and Weyde, 2015];
2. meje med segmenti, ki definirajo melodično ponavljajoče dele, in
3. oceno drsenja višine tonov, ki opisuje globalno spremembo intonacije znotraj

izvedbe.

Metoda združi vse tri vire za oceno višin tonov najbolj reprezentativnega dela skladbe na podlagi vseh ponovitev. Na koncu z verjetnostnim pristopom na podlagi muzikološkega znanja iz višin tonov oceni note ter njihove začetne in končne čase. V nadaljevanju sledi podrobnejša predstavitev posameznih korakov metode.

9.1 Segmentacija in ocena drsenja višine tonov

Segmentacijo in oceno drsenja višine tonov pridobimo z metodo, predstavljeno v prvem delu disertacije. Segmentacijo definiramo kot množico časov:

$$\Psi = \{t_1, t_2, \dots, t_G\}, \tag{9.1}$$

kjer G predstavlja število segmentov v pesmi in t_i začetek i-tega segmenta.

Drsenje tonske višine predstavimo z zaporedjem zamikov tonskih višin od pričakovane intonacije v poltonih za posamezne dele posnetka:

$$\Phi = [\phi_1, \phi_2, \dots, \phi_G], \tag{9.2}$$

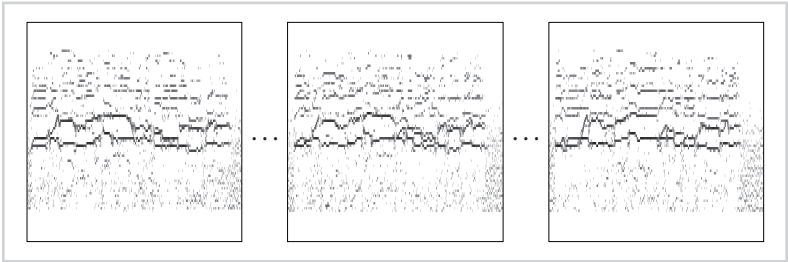
kjer je posamezni del posnetka dolg eno sekundo.

Predstavljena metoda doseže na ljudski glasbi vrednost mere F1 0,76, še bolje pa se odreže prav na večglasni vokalni ljudski glasbi, kjer doseže vrednosti mere F1 0,84 na dvo- in triglasnem petju ter 0,78 na posnetkih zborovske glasbe.

9.2 Ocena višin osnovnih tonov

Za oceno višin osnovnih tonov v zvočnem signalu se zanašamo na obstoječe metode, kot sta na primer Klapuri [Klapuri, 2003] in Benetos [Benetos and Weyde, 2015]. Od metode pričakujemo, da za vsak okvir signala oceni višine prisotnih osnovnih tonov $F = [f_{it}]$ in njihove izrazitosti (angl. salience values) $S = [s_{it}]$, kjer velja $i \in [1, M]$ in $t \in [1, T]$. M predstavlja največje število sočasnih ocen višin osnovnih tonov, T pa

dolžino analiziranega signala. Na sliki 9.2 so prikazane ocene višin osnovnih tonov z metodo Klapuri za tri segmente izbrane pesmi. Temnejša barva v sliki predstavlja tone z višjimi ocenami izrazitosti.



Slika 9.2: Ocene višin osnovnih tonov z metodo Klapuri za tri segmente izbrane pesmi.

9.3 Kompenzacija drsenja višine tonov

Da lahko medsebojno smiselno primerjamo posamezne dele pesmi, moramo popraviti ocenjene višine osnovnih tonov glede na spremembo intonacije, do katere prihaja zaradi drsenja v višini tonov skozi celotno izvedbo. Ker je ocena drsenja D višine tonov že del segmentacije, v tem koraku popravimo ocene višin osnovnih tonov z ocenami drsenja višine tonov kot: $f_{it}^{'} = f_{it} + d_t$, kjer so vse vrednosti podane v centih.

9.4 Izbor reprezentativnega segmenta

Cilj predstavljene transkripcijske metode je pridobiti najbolj reprezentativno transkripcijo melodično ponavljajočega segmenta v pesmi. Ker se vse ponovitve segmentov med seboj razlikujejo zaradi narave izvedbe, je potrebno izbrati segment, ki je reprezentativen za celotno pesem. Takšen segment definiramo kot segment, ki je čimbolj podoben ostalim segmentom v pesmi. Povedano drugače: iščemo takšen segment, ki ni bistveno drugačen od vseh ostalih ponovitev, kar posledično pomeni, da ne vsebuje veliko netočnosti v izvedbi. Reprezentativni segment metoda v nadaljnjih korakih uporabi, da z njim poravna vse ostale segmente in izračuna najbolj verjetno (največkrat odpeto) zaporedje višin tonov v pesmi in posledično najbolj reprezentativno transkripcijo pesmi.

Za iskanje najbolj reprezentativnega segmenta bi lahko uporabili izčrpno poravnavo in primerjavo vseh segmentov, a zaradi časovne zahtevnosti in dejstva, da ne želimo izbrati zgolj izstopajočega segmenta, postopek poenostavimo. Najprej predstavitev frekvenc in ocen izrazitosti, popravljenih s kompenzacijo drsenja višine tonov, pretvorimo v predstavitev v obliki notnega traku oz. predstvitev $P = [p_{it}]$, kjer velja $i \in [1, F_{max}]$ in $t \in [1, T]$ in i predstavlja frekvenčni razred na kvantizirani logaritmični lestvici frekvenc:

$$i = \frac{f}{100}, \tag{9.3}$$

kjer za izhodiščno višino tona $f = 0$ centov privzamemo 27,5 Hz.

Predstavitev P vsebuje nenegativne vrednosti ocen izrazitosti v celicah, ki ustrezajo višinam najdenih tonov. Pridobljeno predstavitev popravimo z oceno drsenja višine tonov.

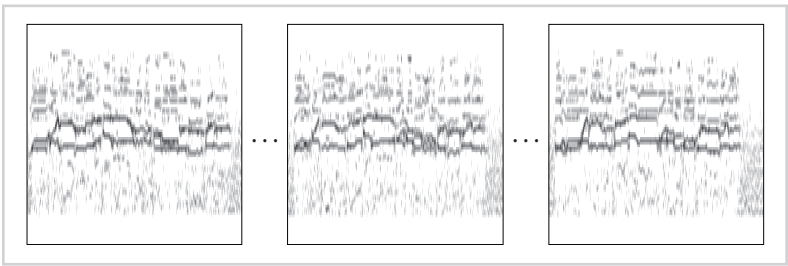
Posamezne segmente P_i izračunane predstavitve $P = [P_1, P_2, \dots, P_H]$, kjer H predstavlja število segmentov, prevzorčimo na velikost najkrajšega segmenta in med njimi izračunamo mero podobnosti z uporabo kosinusne razdalje. Za reprezentativnega izberemo tisti segment, ki doseže najnižjo vrednost kosinusne razdalje do vseh ostalih segmentov.

Predstavljen poenostavljen postopek izbire reprezentativnega segmenta nam zagotavlja, da je izbrani segment po ocenah višin tonov in dolžini podoben ostalim in tako primeren za poravnavo.

9.5 Poravnava segmentov in izračun povzetka

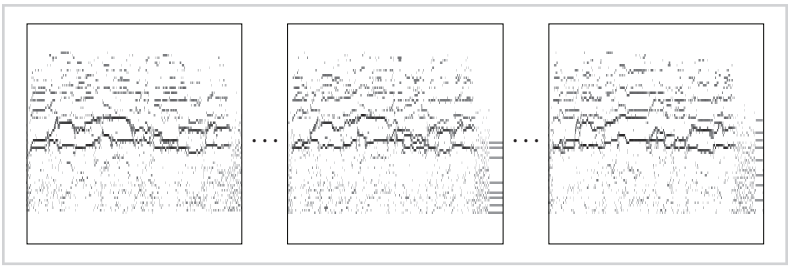
Ker posamezni segmenti s pričetki v časih t_i niso enake dolžine zaradi sprememb v tempu med izvedbo, je potrebno vse segmente v časovni domeni poravnati na izbran reprezentativni segment s pričetkom v t_r. Poravnavo izvedemo z dinamičnim ukrivljanjem časa (DTW) na posameznih parih segmentov, predstavljenih z izvlečki iz predstavitve P_{t_i}. Za mero razdalje uporabimo korelacijsko razdaljo, za katero smo v prvem delu pokazali, da je primerna za ta namen.

Težavo pri poravnavi segmentov predstavljajo lokalne netočnosti in artikulacije v izvedbi, kot so napačna izvedba posamičnih not, vibrato in sprememba višine tonov ob začetku in koncu posameznega tona. Zato pred poravnavo zgladimo predstavitev segmentov P z uporabo Gaussovega filtra preko frekvenčnih kanalov. Izbrali smo filter dolžine 3 s parametrom $\sigma = 1$, s čimer zabrišemo omenjene lokalne netočnosti v obsegu enega poltona in izračun razdalje naredimo robustnejši, poravnavo pa natančnejšo. Na sliki 9.3 je prikazano glajenje predstavitve segmentov preko frekvenčnih kanalov za iste segmente kot na sliki 9.2.



Slika 9.3: Prikaz glajenja za tri segmente izbrane pesmi.

Poravnavo izvedemo s postopkom dinamičnega ukrivljanja časa (DTW), ki smo ga opisali že v prejšnjih poglavjih. Dele signala s tišino v tem primeru nadomestimo z ničelnimi vrednostmi in s tem poskrbimo, da se tudi območja tišine med seboj ustrezno poravnajo. Rezultat poravnave je serija optimalnih poti poravnav med segmentom s pričetkom v t_r in vsemi ostalimi segmenti s pričetki v t_i: $\{\rho_i : i = 1 \ldots |\Psi|\}$. Na sliki 9.4 je prikazana poravnava istih izbranih segmentov kot na sliki 9.2.



Slika 9.4: Prikaz poravnave z DTW za tri segmente izbrane pesmi.

V naslednjem koraku algoritem povzame informacije o višinah osnovnih tonov in njihovih izrazitosti preko vseh segmentov in izračuna robustnejšo oceno višin tonov v

izbranem segmentu. Tukaj sledimo premisleku, da ponovitve prispevajo dodatno informacijo o izvedbi, in tako obravnavamo *povprečno* izvedbo, kjer imajo večkrat ponovljene višine tonov višjo oceno izrazitosti. Rezultat združitve preko vseh segmentov je prikazan na sliki 9.5 (a), kjer so med drugimi tudi segmenti s slike 9.2.

Postopek je sledeč. Definiramo $\mathcal{F}^a$ in $\mathcal{S}^a$ kot množici ocen višin osnovnih tonov in njihovih izrazitosti vseh segmentov ω_j, poravnanih na ω_r glede na pot ρ_j. V vsakem časovnem okviru t združimo pare višin osnovnih tonov in ocen izrazitosti preko vseh segmentov z uporabo požrešne metode, kjer v vsakem koraku tonsko višino z najvišjo oceno izrazitosti (f_t^{max}) in združimo vrednosti preko vseh segmentov:

$$\mathcal{F}_t^{max} = [f_t^{max} - \eta, f_t^{max} + \eta] \tag{9.4}$$

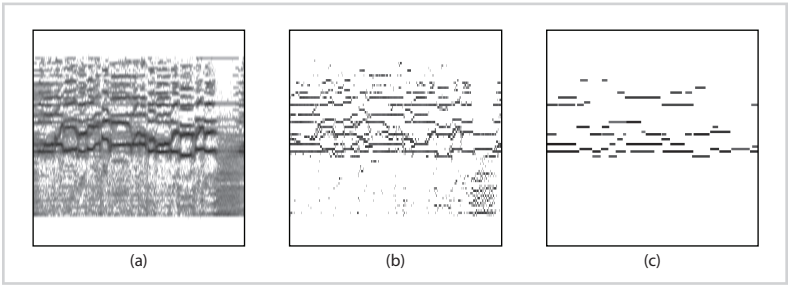
$$s_t^c = \sum_i s_{it}^a \big|_{s_{it}^a \hat{=} f_{it}^a \in \mathcal{F}_t^{max}} \tag{9.5}$$

$$f_t^c = \frac{1}{s_t^c} \sum_i f_{it}^a s_{it}^a \big|_{s_{it}^a \hat{=} f_{it}^a \in \mathcal{F}_t^{max}} \tag{9.6}$$

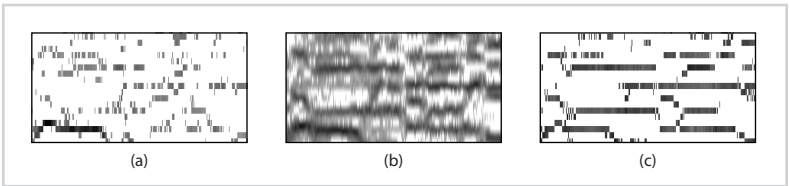
Parameter η definira frekvenčno okno okoli f_t^{max}, za katerega predpostavimo, da vsebovane frekvence predstavljajo isti ton. Zaradi nepopolne izvedbe višina osnovnega tona namreč ne bo sovpadala točno s frekvencami idealno uglašenih tonov in s tem zagotovimo toleranco na tovrstna odstopanja.

Vrednosti f_t^c in s_t^c dodamo v množico vseh povzetih vrednosti $\mathcal{F}^c$ in $\mathcal{S}^c$, vrednosti, uporabljene pri izračunu, odstranimo iz $\mathcal{F}^a$ in $\mathcal{S}^a$ ter postopek ponavljamo, dokler ne preostane 30 % najnižjih ocen izrazitosti, kar smo ocenili glede na rezultate Klapurijevega algoritma, ki ima nizko vrednost mere natančnosti.

Kot rezultat za vsak časovni okvir t dobimo množico ocen višin osnovnih tonov $\mathcal{F}^c$ in njihovih izrazitosti $\mathcal{S}^c$. Množice preko vseh časovnih okvirjev predstavljajo transkripcijo *povprečne* izvedbe skladbe, za katero lahko trdimo, da je najbolj reprezentativna, saj predstavlja najbolj tipično izvedbo skladbe preko vseh segmentov. Ponovitve segmentov prispevajo k stabilnejšim ocenam višin osnovnih tonov z manj napakami, kar v nadaljevanju tudi ovrednotimo. Na sliki 9.5 (b) je prikazan primer povzetka segmentov. Podrobnejši prikaz delovanja metode do povzetka segmentov je prikazan na sliki 9.6.



Slika 9.5: Združeni segmenti (a), povzetek segmentov (b) in končna transkripcija (c).



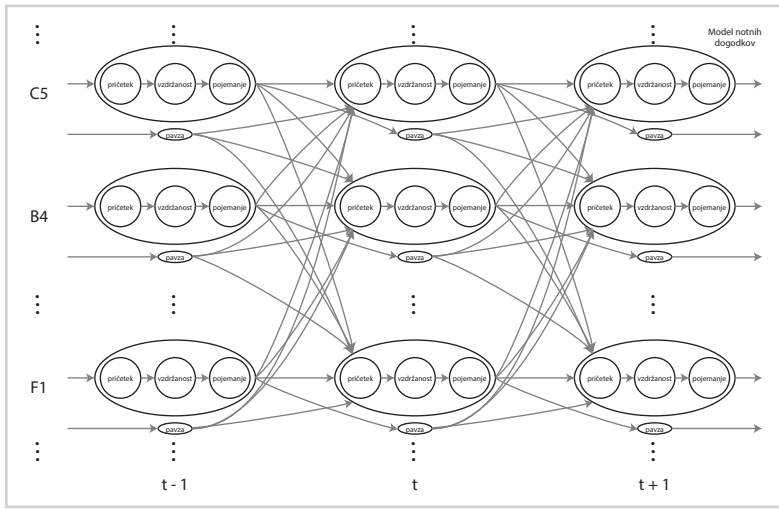
Slika 9.6: En segment (a), združeni segmenti (b), povzetek (c).

9.6 Izračun not

Za izračun končne transkripcije, ki zajema višine ter začetne in končne čase not, moramo zaporedje ocen višin osnovnih tonov in njihovih izrazitosti pretvoriti v zaporedje not. V ta namen smo razvili verjetnostni model, zasnovan na modelu, predstavljenem v [Ryynänen and Klapuri, 2004, 2005], katerega ideja temelji na modelu povezanih besed s področja prepoznave govora [Young et al., 1989]. Model, skiciran na sliki 9.7, temelji na treh podmodelih:

1. *model notnih dogodkov* za modeliranje posameznih not,
2. *model pavz* za modeliranje tišine in pavz ter
3. *muzikološki model*, ki povezuje oba modela in modelira prehode med posameznimi notami in pavzami.

V modelu je vsaka nota, ki jo želimo transkribirati, predstavljena z dvema podmodeloma: notnim modelom, ki predstavlja evolucijo note v vhodnem signalu, in modelom pavze. Transkripcijo izračunamo s ponavljanjem izračuna najbolj verjetnega zaporedja stanj modela preko celotne dolžine signala in odstranitvijo najdenega zaporedja iz



Slika 9.7: Model za oceno not.

signala, kar je podrobneje predstavljeno v nadaljevanju.

9.6.1 Model notnih dogodkov

Posamezno noto modeliramo s skritim markovskim modelom (HMM) s tremi stanji, kjer posamezna stanja predstavljajo pričetek, vzdržanost in pojenjanje note. Za razliko od pristopa [Ryynänen and Klapuri, 2005] in podobno, kot je to izvedeno v bolj aktualnih metodah (npr. [Benetos and Weyde, 2015]), za modeliranje not uporabimo izvedenko modela HMM - HMM z eksplicitno določenimi trajanji (angl. explicit duration hidden Markov model - EDHMM) [Mitchell et al., 1995].

Skriti markovski model z eksplicitno določenimi trajanji obiskov stanj

Skriti markovski model z eksplicitno določenimi trajanji obiskov posameznih stanj je izvedenka običajnega skritega markovskega modela (glej poglavje 4.4), pri katerem eksplicitno modeliramo trajanja obiska posameznega stanja z verjetnostno porazdelitvijo. Ostale definicije ostanejo enake kot pri običajnem HMM-ju.

Prednosti uporabe tako definiranega modela so predvsem v tem, da verjetnostna porazdelitev ostajanja v stanju ni geometrična (kot pri HMM s povratnimi zankami), ampak lahko za modeliranje verjetnosti, da se v stanju nahajamo t časa, uporabimo poljubno verjetnostno porazdelitev, kar je v glasbi še posebej smiselno.

Formalno EDHMM definiramo s šesterko (S, V, T, Π, A, B), kjer je $S = \{s_1, \dots, s_n\}$ množica N stanj sistema, $V = \{v_1, \dots, v_M\}$ množica M izhodnih simbolov, $T = \{\tau_1, \dots, \tau_N\}$ množica verjetnostnih porazdelitev za definicijo trajanja obiska posameznega stanja sistema, $\Pi = \{\pi_1, \dots, \pi_N\}$ množica začetnih verjetnosti stanj, $A = \{a_{ij}\}$ matrika verjetnosti prehodov med stanji sistema in $B = \{b_{i,v_k}\}$ matrika verjetnosti oddajanja simbolov sistema. Z množico $\Lambda = \{T, \Pi, A, B\}$ predstavimo parametre sistema:

- τ_i - prestavlja verjetnostno porazdelitev, ki definira trajanje obiska i-tega stanja sistema;
- π_i - predstavlja verjetnost, da sistem prične v stanju i;
- a_{ij} - predstavlja verjetnost prehoda sistema iz stanja i v stanje j, kjer so prehodi v isto stanje prepovedani ($i \neq j$), in
- b_{i,v_k} - predstavlja verjetnost, da sistem v stanju i odda simbol v_k.

Veljati mora tudi:

$$\sum_{i=1}^{N} \pi_i = 1,$$

$$\sum_{j=1}^{N} a_{ij} = 1;\ i \in [1,\ N],$$

$$\sum_{k=1}^{M} b_{i,v_k} = 1;\ i \in [1,\ N].$$

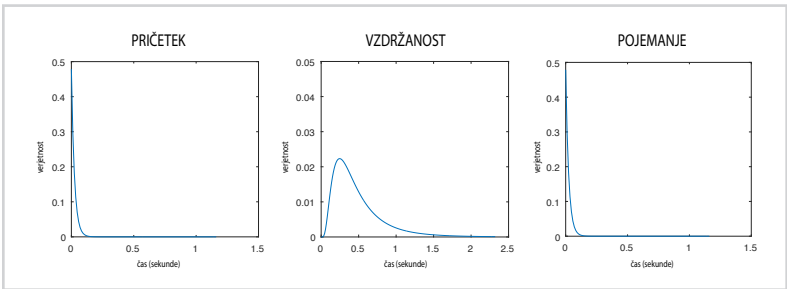
Več o modelih EDHMM je predstavljeno v [Mitchell et al., 1995].

–

Celoten transkripcijski model je sestavljen iz enega modela notnih dogodkov za vsako noto (višino tona), ki jo želimo transkribirati. Število modelov določimo dinamično glede na razpon zaznanih višin osnovnih tonov v predhodnih korakih.

Parametre modela EDHMM določimo, kot je predstavljeno v nadaljevanju. Začetno stanje je vedno stanje pričetka note in končno stanje vedno stanje pojenjanja note. Prehodi so trivialni, saj dopuščamo samo prehode naprej in so posledično verjetnosti prehodov enake 1. Model EDHMM ne dopušča prehoda iz stanja nazaj v isto stanje.

Verjetnost zasedanja posameznega stanja sistema $d_j(u) = P(S_{t+u+1} \neq j, S_{t+2}^{t+u} = j | S_{t+1} = j, S_t \neq j)$ je za stanja pričetkov in pojenjanj definirano z geometrijsko porazdelitvijo (kot je to v običajnih modelih HMM), ki teži h kratkim obiskom posameznih stanj. $d_j(u)$ izhaja iz verjetnostnih porazdelitev trajanja obiska posameznega stanja, definiranih z množico T. Za stanje vzdržanosti uporabimo logaritmično normalno verjetnostno porazdelitev (angl. log-normal distribution), kot jo predlagajo avtorji [Takeda et al., 2007], ki ne preferira zelo kratkih obiskov stanj, po drugi strani pa dopušča daljša trajanja. Porazdelitve so prikazane na sliki 9.8 Parametri vseh porazdelitev so bili izbrani ročno v skladu z muzikološkim znanjem.



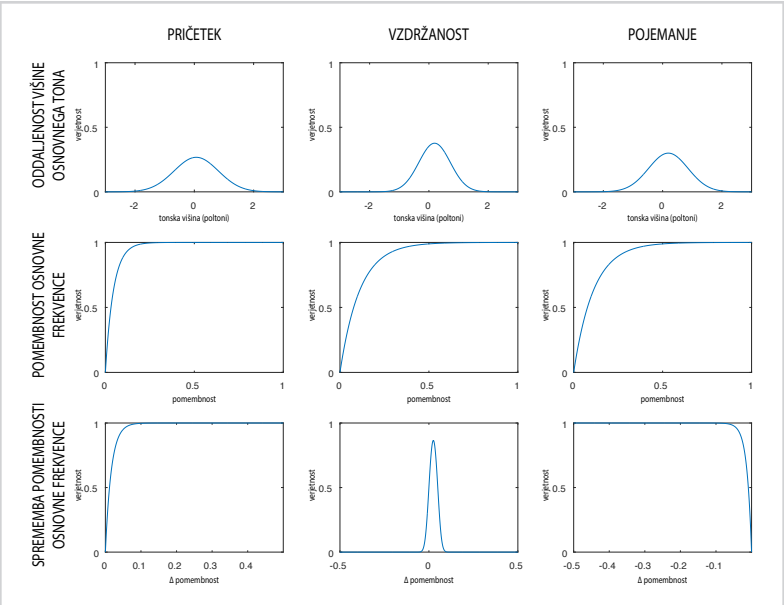
Slika 9.8: Verjetnostne porazdelitve zasedenosti posameznega stanja.

Izhodne emisijske verjetnosti modela temeljijo na treh značilnicah:

- oddaljenost višine osnovnega tona od višine pri idealni uglasitvi;
- izrazitost višine osnovnega tona in
- razlike v vrednostih izrazitosti, ki modelirajo dinamiko.

Vsako značilnico modeliramo z ločeno verjetnostno porazdelitvijo za posamezna stanja

modela, kar skupno predstavlja 9 verjetnostnih porazdelitev, prikazanih na sliki 9.9. Parametre porazdelitev smo ovrednotili na validacijski množici.



Slika 9.9: Verjetnostne porazdelitve, uporabljene za modeliranje verjetnosti oddajanja simbolov posameznega stanja modela notnih dogodkov (pričetek, vzdržanost, pojemanje) za izbrane značilnice.

Oddaljenost višine osnovnega tona od idealne frekvence note modeliramo z normalno porazdelitvijo in ima višjo toleranco v stanjih pričetka in pojemanja ter nižjo v stanju vzdržanosti. To še posebej velja za vokalno glasbo, kjer se lahko višina tonov precej spreminja med pričetkom in zaključkom note, medtem ko v stanju vzdržanosti ostaja relativno stabilna (kljub vsemu moramo upoštevati nekaj tolerance zaradi vibrata in netočnega petja). Za oceno izrazitosti velja, da je nižja ob začetkih not, kjer se višina tonov precej spreminja, in je relativno stabilna v stanju vzdržanosti in pojemanja, zaradi česar jo modeliramo z eksponentno porazdelitvijo. Za dinamiko - spremembo ocene izrazitosti - pričakujemo, da je pozitivna ob začetku, negativna ob pojemanju in se giblje okoli ničle v stanju vzdržanja, kar se prav tako odraža v izbranih porazdelitvah.

Končne izhodne emisijske verjetnosti posameznega stanja modeliramo z uteženim zmnožkom posameznih verjetnosti dogodkov glede na vrednosti značilnic. Uteži pri zmnožku

posameznih verjetnostih dogodkov so konstantne in so izbrane tako, da odražajo izrazitost posamezne značilnice. S primerno izbiro uteži damo več poudarka na oddaljenost od idealne frekvence in manj obema ostalima značilnicama. Posamezne verjetnostne porazdelitve utežujemo zaradi bolj realističnega modeliranja razvoja note skozi čas.

9.6.2 Model pavz

Model pavz modelira dele signala, kjer je verjetnost not nizka. Model definiramo z modelom HMM z enim stanjem. Za razliko od pristopa [Ryynänen and Klapuri, 2005] uporabimo en model pavz za posamezni model notnih dogodkov, kar omogoča fleksibilnost pri ohranjanju melodičnega konteksta preko pavz v melodični liniji in je razvidno tudi na sliki 9.7. Opazovana verjetnost v času t je definirana kot:

$$P(silence)_t = 1 - \max_{i,k}\{P(b_{i,v_k})\}, \tag{9.7}$$

kjer je maksimalna vrednost b_{i,v_k} enaka 1. Povedano drugače, če je opazovana verjetnost v modelu not visoka, potem je verjetnost v modelu pavze nizka in obratno.

9.6.3 Muzikološki model

Muzikološki model določa verjetnosti prehodov med modeli notnih dogodkov in pavz. Za določanje teh verjetnosti uporabimo podoben pristop kot [Ryynänen and Klapuri, 2004, 2005], kjer so verjetnosti prehodov izračunane na zbirki melodij ljudskih pesmi. Posledično verjetnosti kodirajo glasbeno predznanje o tem, kateri prehodi med notami so najbolj verjetni. Ker so verjetnosti prehodov odvisne od glasbenega ključa izvedbe, v nadaljevanju najprej ocenimo glasbeni ključ, nato glede na ključ določimo verjetnosti prehodov med posameznimi notnimi modeli in nazadnje izvedemo transkripcijo.

Ocena glasbenega ključa pesmi

Glasbeni ključ je definiran z osnovno notno lestvico izvedbe pesmi. Prva nota v lestvici je tonika. V našem primeru imamo tonični noti $k_{maj}, k_{min} \in \{0, 1, \ldots, 11\}$, kjer vrednosti $0, 1, \ldots, 11$ odražajo tonske razrede C, C#, …, H. V kolikor durovksa (angl.

major) in molovska (angl. minor) lestvica vsebujeta iste tone, predpostavimo, da sta lestvici relativni, in posledično definiramo relativni par ključev, za katerega velja, da tonične note sledijo pravilu:

$$k_{maj} = mod(k_{min} + 3, 12) \Leftrightarrow k_{min} = mod(k_{maj} + 9, 12), \tag{9.8}$$

kjer *mod* predstavlja ostanek pri deljenju. Za oceno glasbenega ključa iz ocen osnovnih tonov izračunamo profilnico tonskih razredov $h = [h_1, h_2, \dots, h_{12}]$, kjer velja:

$$h_i = \sum_{f_j} mod(round(\frac{f_j}{100}) - 39, 12) + 1. \tag{9.9}$$

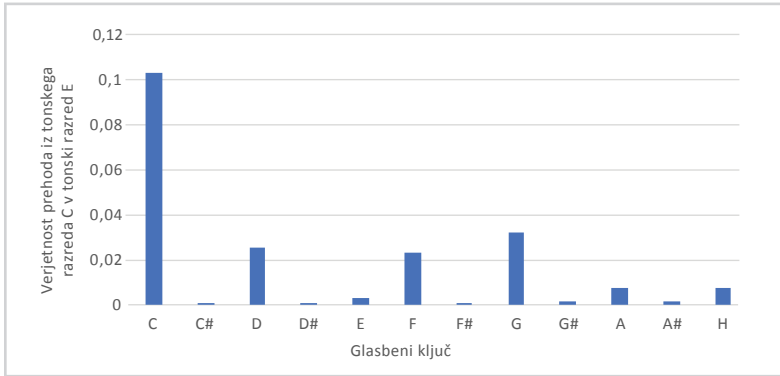
Pridobljeno profilnico h s pomočjo korelacije primerjamo z znanimi profili Krumhanslove [Krumhansl, 1990] in izberemo tista dva, ki najbolje ustrezata durovskemu in pripadajočemu molovskemu načinu.

Izračun verjetnosti prehodov med notnimi modeli

Verjetnosti notnih prehodov so ocenjene s štetjem pojavitev različnih notnih bigramov (intervalov med notami) v zbirki Essen Folksong Database [Schaffrath, 1995], iz katere je bilo uporabljenih več kot pol milijona zaporedij. Predpostavljamo, da so zakonitosti prehodov iz te zbirke, ki vsebuje zapise evropskih ljudskih pesmi, dovolj splošne tudi v našem kontekstu. Glede na ključ izvedbe so verjetnosti prehodov med notnimi modeli sorazmerne s pogostostjo pojavitve intervalov (skokov) med zaporednima notama.

Muzikološki model predpostavlja, da je bolj verjetno pričeti in končati notno zaporedje z noto, ki se v nekem glasbenem ključu pogosto pojavlja, kar uporabimo pri določitvi začetnih verjetnosti stanj. Za primer na sliki 9.10 podajamo graf verjetnosti prehodov iz tonskega razreda C v tonski razred E v različnih glasbenih ključih.

Ker za vsak model notnih dogodkov uporabljamo po en model pavze, so verjetnosti prehodov nota-pavza in pavza-nota enake kot ustrezne verjetnosti prehodov nota-nota. Prehodi pavza-pavza niso dopuščeni.



Slika 9.10: Verjetnosti prehodov iz tonskega razreda C v tonski razred E v različnih glasbenih ključih.

Končna transkripcija

Končno polifonično transkripcijo izračunamo z iskanjem optimalnih poti skozi mrežo predstavljenega dvonivojskega verjetnostnega modela. Vsaka izmed poti določa posamezno melodično linijo transkripcije. Za izračun nabora optimalnih poti skozi mrežo modelov uporabimo algoritem pošiljanja sporočil (angl. message passing algorithm) [Young et al., 1989], podobno kot je predstavljeno v [Ryynänen and Klapuri, 2005], s to razliko, da smo v našem primeru uporabili prilagojen algoritem, ki deluje tudi z uporabo modelov EDHMM.

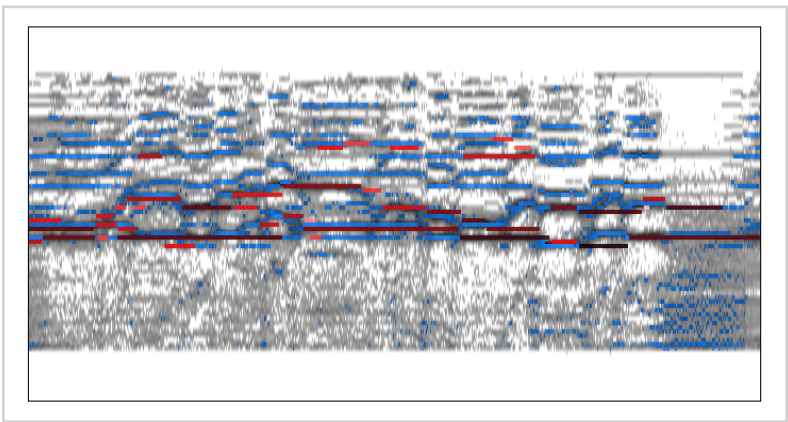
Algoritem deluje tako, da skozi mrežo pošilja posamezne žetone. Vsak obisk modela notnih dogodkov poveča utež sporočilu v skladu z verjetnostmi oddaje simbolov, z verjetnostjo prehodov med stanji in z verjetnostjo ostajanja v stanju. Ko sporočilo zapusti sistem notnih dogodkov, dobimo mejo note, ki jo dodamo na seznam vseh notnih meja za kasnejšo uporabo z algoritmom vračanja (angl. backtracking algorithm). Muzikološki model poveča utež sporočilu na prehodih med posameznimi modeli notnih dogodkov glede na to, katerega izmed modelov notnih dogodkov je sporočilo predhodno zapustilo.

Da najdemo več poti, moramo algoritem pognati večkrat, pri tem pa moramo paziti, da pri ponovnih obhodih izključimo obiske istih modelov notnih dogodkov, kot jih je model že obiskal v predhodnih izvajanjih. S tem zagotovimo, da se melodične linije med seboj ne prekrivajo. Po vsakem zagonu algoritma tako iz mreže modela izključimo

tiste modele notnih dogodkov, ki smo jih v zadnjem zagonu algoritma že obiskali.

Vsakokratna optimalna pot skozi mrežo modela je določena s potjo sporočila z najmanjšo utežjo, ustrezno notno zaporedje pa pridobimo z uporabo algoritma vračanja nad seznamom notnih meja. Algoritem lahko poganjamo tolikokrat, da dosežemo želeno stopnjo polifonije, ali toliko časa, dokler sistem ne vrne zaporedja obiskov modelov pavz.

Primer izračuna not - transkripcije - je prikazan na sliki 9.5 (c). Na sliki 9.11 je prikazana neposredna primerjava rezultatov po združevanju segmentov (siva barva), po izračunu povzetkov združenih segmentov (modra barva) in po končni transkripciji (rdeča barva). Dobro je razvidno, kako posamezni koraki vodijo do *očiščene* končne transkripcije.



Slika 9.11: Združeni segmenti (siva), povzetek segmentov (modra) in končna transkripcija (rdeča).

Ovrednotenje razvite metode

10

Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted.

– Albert Einstein

V poglavju predstavimo ovrednotenje metode za transkripcijo ljudskih pesmi, predstavljene v prejšnjem poglavju.

10.1 Rezultati

V predhodnjem poglavju predstavljeno metodo smo ovrednotili na zbirki 37-ih vokalnih večglasnih ljudskih pesmi. Zbirka obsega 107,7 minut zvočnih posnetkov, ki so del terenskih posnetkov ljudske glasbe iz arhiva EtnoMuza [Strle and Marolt, 2007]. Transkripcije vseh pesmi so bile izdelane ročno s strani etnomuzikologov in so bile v časovni domeni polavtomatsko poravnane z zvočnimi posnetki. Pesmi v zbirki so bile izbrane tako, da zajemajo različne vidike problemov, na katere naletimo v terenskih posnetkih ljudske glasbe, kot so netočno petje, drsenje višine tonov in slaba kvaliteta posnetkov. Prav tako smo ovrednotili tudi oba glavna koraka predlagane metode: (1) poravnava in povzetek osnovnih frekvenc in pomembnosti in (2) izračun not.

10.1.1 Ovrednotenje trenutno aktualnih transkripcijskih metod

Za primerjavo uspešnosti metode smo predlagano transkripcijsko metodo primerjali z že obstoječimi pristopi za glasbeno transkripcijo. Pri tem smo izbrali pristope: Klapuri [Klapuri, 2003], Sonic [Marolt, 2004], Silvet [Benetos and Dixon, 2013] in Benetos-2015 [Benetos and Weyde, 2015]. Ker pristopa Klapuri in Benetos-2015 vračata ocene osnovnih frekvenc in pomembnosti za posamezni okvir signala, smo izhoda teh metod uporabili kot vhod za predlagani pristop. Evalvacijo smo izvedli na časovni ločljivosti 40 ms s frekvenčno toleranco odstopanja četrt tona. Poleg točne transkripcije smo ločeno ovrednotili tudi primer, ko zanemarimo oktavne napake, ki so v določenih primerih povsem sprejemljive. Vrednosti mere F1 so predstavljene v tabeli 10.1.

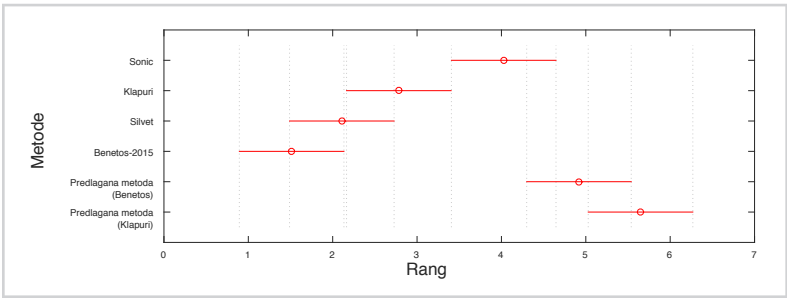
10.1.2 Ovrednotenje predlagane metode

Rezultati ovrednotenja predlaganega pristopa kažejo, da signifikantno izboljša rezultate obeh uporabljenih osnovnih metod (več kot 50 % izboljšanje vrednosti mere F1). Izboljšanje je večje za natančnost (ta se skoraj podvoji), a je tudi izboljšanje priklica zadovoljivo (okoli 15 %). Rezultati s Klapurijevim algoritmom so nekoliko boljši, vendar v kolikor ignoriramo oktavne napake, je vrednost mere F1 v obeh primerih 0,59,

Tabela 10.1: Rezultati ovrednotenja aktualnih transkripcijskih metod in predlaganega pristopa na zbirki ljudskih pesmi. Mere natančnost (P), priklic (R) in F1 so izračunane kot povprečna vrednost mer za posamezno pesem

Metoda	P	R	F1
Sonic	0,39	0,50	0,43
Klapuri	0,26	0,59	0,36
Silvet	0,25	0,46	0,32
Benetos-2015	0,21	0,47	0,29
predlagana metoda (Klapuri)	0,50	0,68	0,58
predlagana metoda (Benetos)	0,51	0,55	0,52
predlagana metoda (Klapuri) - brez oktavnih napak	0,52	0,71	0,59
predlagana metoda (Benetos) - brez oktavnih napak	0,58	0,62	0,59

kar pomeni, da pristop Benetos-2015 že v osnovi generira več oktavnih napak. Kot je razvidno iz rezultatov, dosega predlagana metoda z Benetosovim algoritmom precej uravnoteženi vrednosti mer natančnosti in priklica, medtem ko ima Klapurijev algoritem brez upoštevanja oktavnih napak bistveno višjo vrednost priklica 0,71 na račun nekoliko nižje vrednosti mere natančnosti 0,52.



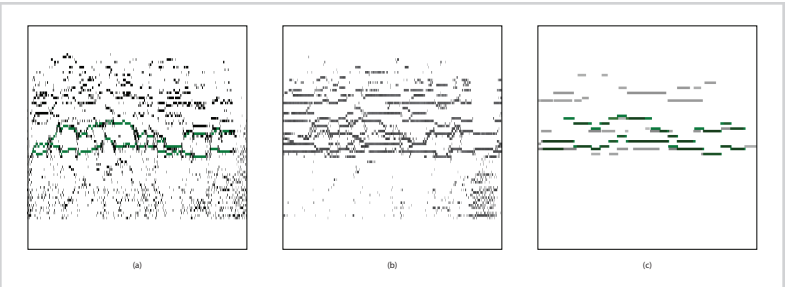
Slika 10.1: Statistična primerjava transkripcijskih metod s Friedmanovim testom.

Izvedli smo tudi statistično primerjavo metod z uporabo Friedmanovega statističnega testa. Rezultati so predstavljeni na sliki 10.1. S slike je razvidno, da je predlagana metoda (Klapuri) signifikantno boljša od vseh obstoječih pristopov, kar pa ne veja za predlagano metodo (Benetos), ki je signifikantno boljša od vseh obstoječih metod, razen metode Sonic.

Analiza delovanja metode

Za analizo delovanja metode smo izbrali primere, na katerih predlagana metoda doseže veliko izboljšanje rezultatov, in po drugi strani primere, kjer izboljšanja ni.

Primer dobrega delovanja s Klapurijevim transkripcijskim algoritmom je prikazan na sliki 10.2, kjer slika 10.2 (a) prikazuje izhod metode Klapuri (siva) in rezultat transkripcije po upragovanju (zelena), slika 10.2 (b) rezultat predstavljene metode po izračunu povzetka informacij iz vseh delov, slika 10.2 (c) pa končni rezultat predstavljene metode (siva) in ročno transkripcijo (zelena).

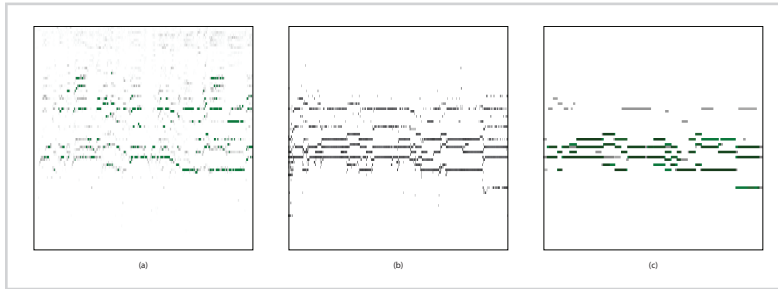


Slika 10.2: Dober primer: (a) Klapuri, (b) povzetek segmentov, (c) rezultat in ročna transkripcija.

S slike je razvidno, da vrne metoda Klapuri kar nekaj kratkotrajnih notnih dogodkov. Rezultat osnovne metode doseže vrednost mere F1 0,22, po združitvi informacij iz ponavljajočih delov pa naša metoda doseže vrednost mere F1 0,33 in s tem 50 % izboljšanje. Končni rezultat naše metode po transkipciji doseže vrednost mere F1 0,52 in s tem osnovni rezultat izboljša za 136 %.

Primer dobrega delovanja s transkripcijskim algoritmom Benetos-2015 je prikazan na sliki 10.3, kjer slika 10.3 (a) prikazuje izhod metode Benetos-2015 (siva) in rezultat transkripcije po upragovanju (zelena), slika 10.3 (b) rezultat predstavljene metode po izračunu povzetka informacij iz vseh delov, slika 10.3 (c) pa končni rezultat predstavljene metode (siva) in ročno transkripcijo (zelena).

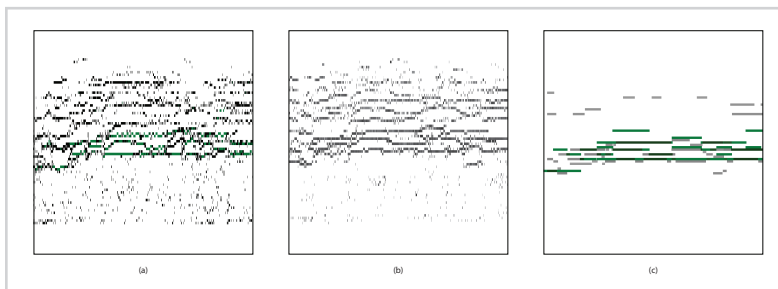
S slike lahko razberemo, da metoda Benetos-2015 zelo razdrobi posamezne notne dogodke. Rezultat osnovne metode doseže vrednost mere F1 0,36. Naša metoda po koraku izračuna povzetka doseže vrednost mere F1 0,53, kar predstalja 47 % izbolj-



Slika 10.3: Dober primer: (a) Benetos, (b) povzetek segmentov, (c) rezultat in ročna transkripcija.

šanje. Končna transkripcija pa doseže vrednost mere F1 0,59, kar predstavlja 63 % izboljšanje.

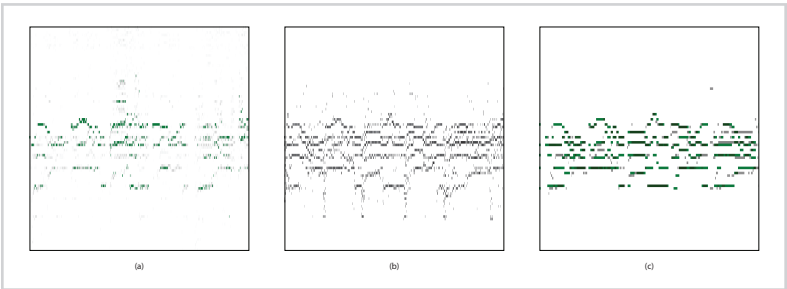
Analizirali smo tudi nekatere primere, na katerih se predlagana metoda najslabše obnese. Primer pri uporabi vhoda metode Klapuri je prikazan na sliki 10.4, kjer slika 10.4 (a) prikazuje izhod metode (siva) in rezultat transkripcije po upragovanju (zelena), slika 10.4 (b) rezultat po izračunu povzetka informacij iz posameznih segmentov, slika 10.4 (c) pa končni rezultat predstavljene metode metode (siva) in ročno transkripcijo (zelena).



Slika 10.4: Slab primer: (a) Klapuri, (b) povzetek segmentov, (c) rezultat in ročna transkripcija.

S slike je razvidno, da metoda Klapuri z upragovanjem odstrani veliko kratkotrajnih notnih dogodkov. Rezultat osnovne metode Klapuri doseže vrednost mere F1 0,48. Naša metoda po koraku izračuna povzetka doseže vrednost mere F1 0,60, kar predstavlja 25 % izboljšanje rezultatov, po končni transkripciji pa naša metoda doseže vrednost mere F1 0,66, kar predstavlja 38 % izboljšanje. Te izboljšave so občutno slabše kot v najboljših primerih.

Primer slabega delovanja z uporabo rezultatov metode Benetos-2015 je prikazan na sliki 10.5, kjer slika 10.5 (a) prikazuje izhod metode (siva) in rezultat transkripcije po upragovanju (zelena), slika 10.5 (b) rezultat po izračunu povzetka vseh segmentov, slika 10.5 (c) pa končni rezultat transkripcije (siva) in ročno transkripcijo (zelena).



Slika 10.5: Slab primer: (a) Benetos, (b) povzetek segmentov, (c) rezultat in ročna transkripcija.

S slike lahko razberemo, da metoda Benetos-2015 zopet vrne precej razdrobljene notne dogodke. Rezultat osnovne metode Benetos-2015 doseže vrednost mere F1 0,16. Naša metoda po koraku izračuna povzetka doseže vrednost mere F1 0,17, kar predstavlja zgolj 6 % izboljšavo. Po končni transkripciji naša metoda doseže vrednost mere F1 0,25, kar predstavlja 56 % izboljšanje rezultatov osnovne metode.

Delovanje metode ni odvisno od stopnje polifonije v posnetkih, rezultati prav tako ne nakazujejo, odvisnosti od drsenja višine tonov. Metoda deluje najslabše na posnetkih, kjer posamezni pevci zgolj poskušajo slediti vodilnemu na drugi melodični liniji, a jim to ne uspeva.

Analiza delovanja posameznih korakov metode

Analiza rezultatov je pokazala tudi, da je izboljšava dosežena z obema glavnima korakoma metode. Prispevka obeh delov sta precej uravnotežena, saj poravnava in povzetek osnovnih frekvenc in pomembnosti prinese 25 % izboljšanje vrednosti mere F1, izračun not pa še dodatno 25 % izboljšanje vrednosti mere F1. To nakazuje, da predlagan pristop zbere veliko informacij iz samih ponovitev, prav tako pa doseže signifikantno izboljšanje tudi z uporabo predlaganega modela za oceno not.

10.2 Ovrednotenje robustnosti metode

Podobno kot za metodo segmentacije, predstavljene v prvem delu, smo tudi transkripcijsko metodo razvili z namenom uporabe na posnetkih ljudske glasbe, za katere velja, da lahko vsebujejo veliko nepravilnosti in šumov, tako zaradi stare ali slabe snemalne opreme kot tudi zaradi slabih snemalnih pogojev. Podobno kot v prvem delu smo ovrednotili tudi robustnost predstavljene transkripcijske metode na različne degradacije zvočnih posnetkov.

Za testiranje degradacij smo kot pri segmentaciji uporabili ogrodje ADT [Mauch and Ewert, 2013], ki omogoča različne načine in stopnje degradacije zvočnih posnetkov.

Robustnost smo testirali z uporabo Klapurijevega transkripcijskega algoritma, saj daje boljše rezultate od Benetosovega. Prav tako smo omejili degradacijo zgolj na transkripcijo, nismo pa je upoštevali pri izračunu segmentacije, saj smo želeli testirati robustnost transkripcije. Kot smo pokazali že v poglavju 6.2, degradacija na segmentacijo sicer ne vpliva bistveno.

Degradacije, za katere smo izvedli ovrednotenje, so sledeče:

- *dodajanje šuma* - kjer smo se odločili za rdeči, roza in beli šum pri razmerjih signal-šum 40, 20, 10 in 0.
- *dodajanje zvoka* - kjer smo se odločili za dodajanje posnetka okolja v lokalu in star zaprašen posnetek in brnenje 50 Hz pri razmerjih signal-šum 40, 30, 20, 10 in 0.
- *rezanje*, kjer smo odrezali najglasnejši del signala, ki zajema 1 %, 2 %, 5 % in 10 % celotnega signala.
- *kompresija*, s pragom kompresije -10, -20, -30 in -40 dB ter stopnjo kompresije 0,2, 0,4, 0,6, 0,8 in 1,0.
- *nizkoprepustni filter* z mejno frekvenco 8.000, 4.000 in 2.000 Hz.
- *visokoprepustni filter* z mejno frekvenco 55, 110, 220 in 440 Hz.
- *harmonična popačenost* z enkratno do petkratno ponovitvijo degradacije popačenosti.

- *stiskanje mp3* z izhodnim podatkovnim tokom 128, 96 in 64 kB/s.

Pri ovrednotenju robustnosti smo uporabljali enak pristop kot pri ovrednotenju segmentacijske metode v poglavju 6.2. Tako v nadaljevanju predstavimo le rezultate, podrobnosti posamezne degradacije pa so navedene v omenjenem poglavju.

10.2.1 Dodajanje šuma

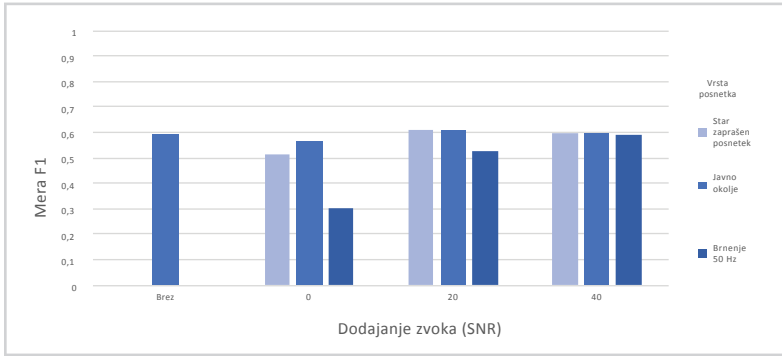
V najslabšem primeru so rezultati po degradaciji slabši za 5 %, v večini primerov pa so rezultati enakovredni tistim brez degradacije. Medtem ko se rezultati osnovne metode z večanjem degradacije slabšajo, predlagana metoda dosega vedno večjo izboljšavo, kar se odraža v enakih končnih rezultatih. Razlog za takšno delovanje je v koraku povprečenja segmentov, kjer zaradi združevanja informacij več segmentov šum izgubi vpliv.

10.2.2 Dodajanje zvoka

Rezultati pri dodajanju zvočnega posnetka v obstoječi signal so precej odvisni od tega, kateri posnetek dodamo v signal in v kakšnem razmerju. V primeru posnetka *star zaprašen posnetek* so v najslabšem primeru rezultati metode slabši za 8 %, kar je večinoma razlog poslabšanja rezultatov osnovne metode. Izboljšanje z našo metodo je večinoma približno enako. V primeru posnetka *posnetek okolja v lokalu* se rezultati osnovne metode slabšajo, izboljšanje z našo metodo pa povečuje, kar končne rezultate ohranja na približno enaki vrednosti. Dodajanje posnetka *50 Hz* pa delovanje tako osnovne metode kot izboljšanje z našo metodo precej poslabša. V najslabšem primeru so končni rezultati slabši za skoraj 30 %, kar je razvidno tudi s slike 10.6. Do takšnega poslabšanja pride zaradi tega, ker je jakost frekvence 50 Hz tako izrazita, da jo naša metoda privzame kot eno izmed melodičnih linij. Posledično metoda zajame eno melodično linijo manj in s tem rezultate poslabša tako zaradi transkribiranih tonov pri 50 Hz kot zaradi zgrešenih tonov v manjkajoči melodični liniji.

10.2.3 Rezanje

Degradacija z efektom rezanja na rezultate metode nima velikega vpliva. V nekaterih primerih so rezultati metode celo malenkost boljši kot brez degradacije, a ne več kotkot



Slika 10.6: Rezultati metode pri degradaciji z dodajanjem zvoka v posnetek.

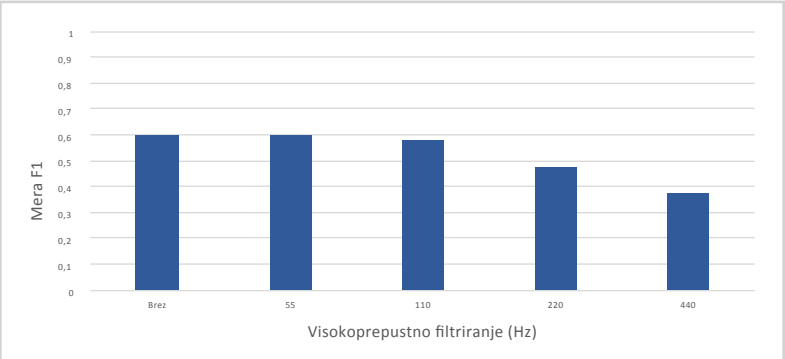
2 %. Podobno kot pri segmentaciji lahko to pojasnimo s tem, da ko iz signala odstranimo najglasnejše dele, ki v večini primerov predstavljajo nenamerne zvočne dogodke, ki v posnetkih predstavljajo zvoke in glasove iz okolice (kriki, zvonenje ure, ploskanje občinstva ipd.).

10.2.4 Kompresija dinamičnega razpona

Degradacija kompresije dinamičnega razpona na delovanje naše, kot tudi na delovanje osnovne, metode ne vpliva, saj so odstopanja v rezultatih znotraj ± 2 %.

10.2.5 Visokoprepustno filtriranje

Z višanjem praga visokoprepustnega filtriranja se rezultati metode občutno poslabšajo, kar je prikazano tudi na sliki 10.7. V najslabšem primeru, ko izvedemo visokoprepustno filtriranje s pragom 440 Hz, se poslabšanje rezultatov odraža v znižanju mere F1 za 0,23 oz. za 38 %. To je povsem razumljivo, saj transkribiramo petje, kjer se do 440 Hz nahaja precejšen del signala, ki predstavlja vokale, kar potrjuje tudi dejstvo, da so rezultati metode pri pragu 55 Hz praktično enaki kot brez degradacije. Poslabša se že delovanje osnovne metode, kar pojasni tudi poslabšanje delovanja predstavljene metode.



Slika 10.7: Rezultati metode pri degradaciji z visokoprepustnim filtriranjem.

10.2.6 Nizkoprepustno filtriranje

Z nižanjem praga nizkoprepustnega filtriranja se rezultati metode minimalno poslabšajo. V najslabšem primeru, ko prag postavimo na 2.000 Hz, se rezultati metode poslabšajo za vsega 1 %, kar je porazdeljeno tako na poslabšanje rezultatov osnovne metode kot na stopnjo izboljšanja, ki ga doseže naša metoda.

10.2.7 Harmonično popačenje

Podobno kot efekt rezanja tudi harmonično popačenje skoraj ne vpliva na rezultate metode. Kljub manjšim spremembam lahko opazimo, da se z večanjem stopnje harmoničnega popačenja rezultati osnovne metode nekoliko poslabšajo (manj kot 2 %), izboljšava predstavljene metode pa nekoliko povečuje (prav tako manj kot 2 %).

10.2.8 Kompresija mp3

Kompresija mp3 ne vpliva niti na rezultate osnovne metode niti na rezultate naše metode. Odstopanja v rezultatih so manj kot 0,2 %, kar je zanemarljivo. Rezultati nakazujejo, da se naša metoda ne zanaša na dele informacij v signalu, ki jih kompresija mp3 odstrani.

10.2.9 Sklep o robustnosti

S tem smo pokazali, da je razvita metoda precej robustna in neobčutljiva na testirane degradacije.

Edina vrsta degradacij, ki močneje vpliva na delovanje metode, je visokoprepustno filtriranje, kjer pa je slabše delovanje razumljivo, saj se ravno v tem delu spektra nahaja največ informacij, potrebnih za uspešno transkripcijo.

Zaključki

II

No one wants to die. Even people who want to go to heaven don't want to die to get there. And yet death is the destination we all share. No one has ever escaped it. And that is as it should be, because Death is very likely the single best invention of Life. It is Life's change agent. It clears out the old to make way for the new.

– Steve Jobs

V drugem delu disertacije smo podrobno predstavili problematiko glasbene transkripcije na domeni ljudske glasbe, kjer splošne transkripcijske metode ne dajejo zadovoljivih rezultatov. Na zbirki večglasne ljudske glasbe, ki predstavlja del arhiva Etnomuza, smo ovrednotili nekaj aktualnih transkripcijskih metod in izpostavili njihove slabosti. Glavne slabosti aktualnih metod so v tem, da so zelo prilagojene za transkripcijo instrumentalne glasbe. Prav tako ne naslavljajo specifik ljudske glasbe, kot so visoka stopnja šuma, slabi snemalni pogoji, amaterski pevci ipd. Po drugi strani pa za ljudsko glasbo velja, da so pesmi v večini primerov sestavljene iz ponovitev istega melodičnega dela.

Na podlagi ugotovitev, pridobljenih iz ovrednotenja aktualnih pristopov, smo razvili lastno transkripcijsko metodo, ki naslavlja in izkorišča lastnosti ljudske glasbe. Posledično so rezultati metode boljši od aktualnih pristopov. V našem primeru smo se usmerili predvsem v transkripcijo večglasne vokalne glasbe, ki aktualnim pristopom predstavlja še poseben izziv.

Razvita metoda kot enega izmed vhodov vzame ocene višin osnovnih tonov, ki jih pridobimo z obstoječimi pristopi. V našem primeru smo za vhode uporabili ocene metod Klapuri [Klapuri, 2003] in Benetos-2015 [Benetos and Weyde, 2015]. Metoda Benetos-2015 je dosegla najboljše rezultate na evalvaciji MIREX 2015. Tako lahko neposredno primerjamo izboljšanje rezultatov s predstavljeno metodo v primerjavi z uporabljenima algoritmoma. Rezultate metode Klapuri izboljšamo za 0,22, rezultate metode Benetos-2015 pa za 0,23. Z vhodom metode Klapuri na glasbeni zbirki doseže vrednost mere F1 0,58. S tem preseže rezultate metode Sonic, ki doseže vrednost mere F1 0,43, za 0,15.

Predstavljeno metodo smo ovrednotili tudi z vidika robustnosti. Rezultati nakazujejo na to, da večina degradacij na samo delovanje ne vpliva bistveno. Največje poslabšanje prinese mešanje signala s signalom posnetka brnenja 50 Hz v razmerju signal-šum 0, ki vrednost mere F1 zmanjša za 0,29. Drugo največje poslabšanje pa je pri uporabi visokoprepustnega filtriranja, kjer z višanjem frekvence odstranimo veliko informacij iz signala, ki se odraža v za 0,23 nižji vrednosti F1.

S predstavljeno metodo smo zadostili zastavljenemu cilju, ki smo si ga zadali v dispoziciji doktorske disertacije z izvirnim znanstvenim prispevkom:

Robusten algoritem za transkripcijo vokalnih ljudskih pesmi, odporen na šume in prekinitve

v posnetkih, ki upošteva netočno petje izvajalcev.

Poleg omenjenega cilja smo s predstavljeno metodo naslovili tudi cilj iskanja najbolj reprezentativnega dela pesmi, saj smo s predstavljeno metodo izvedli transkripcijo reprezentativnega dela, izbranega na podlagi podobnosti z vsemi ostalimi deli celotne pesmi. To se odraža v znanstvenem prispevku:

Na podlagi prejšnjih prispevkov razvit pristop za iskanje transkripcije najbolj reprezentativnega dela.

11.1 Nadaljnje delo

V nadaljevanju podajamo nekaj možnih nadaljnjih korakov:

- Rezultati metode bi lahko bili boljši, če pri transkripciji ne bi upoštevali prav vseh ponovitev segmenta v pesmi. Nekateri segmenti vsebujejo veliko napak izvajalcev ali visoko stopnjo šuma in z njihovim upoštevanjem vsekakor poslabšamo rezultate končne transkripcije. Nekatere ponovitve vsebujejo prekinitve, pri nekaterih prihaja do večjih netočnosti v petju, nekatere pa vsebujejo res visoko stopnjo ozadnega šuma. Če bi uspeli oceniti, katere so res slabe ponovitve, bilahko dodatno izboljšali končno transkripcijo pesmi. Slabe ponovitve bi lahko ocenili na podlagi analize zvočnega posnetka, ali pa bi analizirali samo transkripcijo posameznega dela in jo ovrednotili z muzikološkim modelom.
- Metodo bi lahko razširili tako, da bi za vhod sprejemala ocene osnovnih tonov več metod in bi pri izračunu končne transkripcije to tudi upoštevala. Če bi uporabili metode, ki so bolj prilagojene za posamezni tip glasbe (instrumental, petje ipd.), bi lahko ocene osnovnih tonov primerno utežili glede na tonsko višino posamične melodične linije in s tem poskušali izboljšati rezultate.
- Metodo bi lahko predelali za bolj splošno uporabo in ne zgolj za ljudsko glasbo. Tako bi lahko v popularni ali klasični glasbi iskali ponavljajoče vzorce. V popularni glasbi so takšni vzorci kitice in refreni. Transkripcijo bi lahko s predstavljeno metodo izvedli za vsak nabor ponavljajočih delov. Končno transkripcijo celotne pesmi pa bi sestavili iz posameznih transkribiranih delov.

Sklepna beseda

V pričujoči doktorski disertaciji naslavljamo glasbeno transkripcijo in segmentacijo, dve pomembni področji interdisciplinarnega raziskovalnega področja pridobivanja informacij iz glasbe. Kot je predstavljeno v poglavju 1, kjer podamo motivacijo, se soočamo s problematiko specifične domene - ljudske glasbe, ki zaradi svojih specifik zasluži posebno obravnavo. V uvodu prav tako predstavimo cilje doktorske disertacije in predvidene prispevke znanosti. V nadaljevanju predstavimo pregled področja in izpostavimo probleme obstoječih pristopov na domeni, ki jo naslavljamo.

Predstavljene probleme obravnavamo v dveh delih. V prvem delu naslovimo problem segmentacije ljudske glasbe. Predstavimo evalvacijo obstoječih metod in v nadaljevanju predstavimo segmentacijsko metodo, ki problem segmentacije rešuje bolje od obstoječih metod, kar pokažemo z evalvacijo na zbirki posnetkov ljudske glasbe. Poleg primerjave z ostalimi metodami razvito metodo ovrednotimo tudi s stališča robustnosti delovanja glede na različne degradacije posnetkov. Ne koncu podamo zaključne izsledke in predloge za bodoče izboljšave predstavljene metode.

V drugem delu naslovimo problem transkripcije večglasne ljudske glasbe in iskanja reprezentativnega dela v posnetku. Najprej predstavimo rezultate evalvacije aktualnih transkripcijskih metod na zbirki posnetkov večglasne ljudske glasbe. V nadaljevanju predstavimo lastno transkripcijsko metodo, ki hkrati vrne tudi reprezentativno transkripcijo pesmi. Metodo ovrednotimo in pokažemo, da rešuje problem transkripcije večglasne ljudske glasbe bolje od obstoječih metod. Razvito metodo ovrednotimo tudi s stališča robustnosti. Na koncu podamo sklepne ugotovitve ter predloge za možne izboljšave in prilagoditve predstavljene metode.

Slovarček izrazov

12

Izraz	*Prevod*	*Kratica*
algoritem MCMC	Markov chain Monte Carlo	
algoritem pošiljanja sporočil	message passing algorithm	
algoritem maksimizacije pričakovanja	expectation-maximization algorithm	EM
algoritem vračanja	backtracking algorithm	
analiza neodvisnih komponent	independent component analysis	ICA
človeška analiza zvočne scene	human auditory scene analysis	
dinamično ukrivljanje časa	dynamical time warpping	DTW
DTW za uporabo s podzaporedji	subsequence DTW	
durovksa lestvica	major scale	
Fo ojačane značilnice CENS	Fo enhanced CENS features	
format MIDI	Musical Instrument Digital Interface	MIDI
format WAVE	Waveform Audio File Format	WAVE
glasbeni zvočni povzetek	music audio summary	
harmonične profilnice tonskih razredov	harmonic pitch class profiles	HPCP
hitra Fouriereva transformacija	fast fourier transform	FFT
HMM z eksplicitno določenimi trajanji	explicit duration hidden Markov model	EDHMM
kitica	verse	
kompresija dinamičnega razpona	dynamic range compression	
konstantna Q transformacija	constant Q-transform	CQT
konveksna nenegativna matrična faktorizacija	convex non-negative matrix factorisation	CNMF
kratkočasovna povprečna energija	short-time mean-square power	STMSP
krivulja oddaljenosti	distance curve	
kromatične značilnice	chroma features	
kromatični razredi	pitch classes ali chroma classes	
kromatični vektor	chroma vector	
latentne značilnice strukturnih ponovitev	latent structural repetition features	

Izraz	*Prevod*	*Kratica*
logaritmična normalna verjetnostna porazdelitev	log-normal distribution	
logaritmično zakasnjen korelogram	log-lag correlogram	
matrika časovnih zamikov	time-lag matrix	
Mel-frekvenčni kepstralni koeficienti	Mel-frequency cepstral coefficients	MFCC
metoda analize posamičnih podprostorov	prior subspace analysis	PSA
metoda harmonično adaptivne latentne analize komponent	harmonic adaptive latent component analysis	
metoda izrisa ponovitev	recurrence plot	
metoda k-središč	k-means	
metoda nenegativne matrične faktorizacije	non-negative matrix factorization	NMF
metoda podpornih vektorjev	support vector machines	SVM
metoda nenegativne matrične aproksimacije	non-negative matrix approximation	NNMA
model mešanih Gaussovih verjetnostnih porazdelitev	Gaussian mixture model	GMM
molovska lestvica	minor scale	
na zamik neodvisna metoda PLCA	shift-invariant probabilistic latent component analysis	
nadsegmentacija	oversegmentation	
natančnosti	precision	
ocena izrazitosti	salience value	
ogrodje za analizo glasbene strukture	music structure analysis framework	MSAF
ogrodje za degradacijo zvočnih posnetkov	audio degradation toolbox	ADT
podsegmentacija	undersegmentation	
pojenjanje	decay	
povzemanje glasbe	music summarization oz. audio thumbnailing	
prekrivanje	aliasing	
pričetek	attack	

Izraz	*Prevod*	*Kratica*
pridobivanje informacij iz glasbe	music information retrieval	MIR
priklic	recall	
pristop največjega verjetja	maximum likelihood approach	
profilnice tonskega razreda	pitch class profile	
računski modeli človeškega slušnega sistema	computational models of the human auditory system	
razmerje signal-šum	signal-to-noise ratio	
refren	chorus	
rezanje	clipping	
samopodobnostna matrika	self-similarity matrix	
skriti markovski model	hidden Markov model	HMM
spektralna teorija grafov	spectral graph theory	
spektralno gručenje	spectral clustering	
spremenljiva Q-Transformacija	variable Q-transform	
tonične mreže	tonal centroid features	Tonetz
verjetnostna latentna analiza	probabilistic latent component analysis	PLCA
verjetnostna latentna analiza, neodvisne od zamika	shift-invariant probabilistic latent component analysis	SI-PLCA
višina tona	pitch	
vzdržanost	sustain	
zaporedna linearna diskriminantna analiza	ordinal linear discriminant analysis	
zasičenje	saturation	
značilnice CENS	chroma energy normalized statistics	

LITERATURA

S.A Abdallah. *Towards Music Perception by Redundancy Reduction and Unsupervised Learning in Probabilistic Models*. PhD thesis, King's College London. Department of Electronic & Electrical Engineering, 2002.

Samer A. Abdallah and Mark D. Plumbley. Polyphonic transcription by non-negative sparse coding of power spectra. In *Proceedings of the 5th International Conference on Music Information Retrieval*, Barcelona, Spain, October 2004. URL http://ismir2004.ismir.net/proceedings/p058-page-318-paper216.pdf.

Mark A. Bartsch and Gregory H. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *Applications of Signal Processing to Audio and Acoustics*, pages 15–19, New Platz, NY , USA, October 2001.

Mark A. Bartsch and Gregory H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *Multimedia, IEEE Transactions on*, 7(1):96–104, Feb 2005. ISSN 1520-9210. doi: 10.1109/TMM.2004.840597.

Mert Bay, Andreas F. Ehmann, and J. Stephen Downie. Evaluation of multiple-f0 estimation and tracking systems. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 315–320, Kobe, Japan, October 2009. URL http://ismir2009.ismir.net/proceedings/PS2-21.pdf.

Mert Bay, Andreas F. Ehmann, James W. Beauchamp, Paris Smaragdis, and J. Stephen Downie. Second fiddle is important too: Pitch tracking individual voices in polyphonic music. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, Porto, Portugal, October 2012. URL http://ismir2012.ismir.net/event/papers/319-ismir-2012.pdf.

Emmanouil Benetos and Simon Dixon. Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. *Journal of the Acoustical Society of America (JASA)*, 133(3):1727–1741, 2013.

Emmanouil Benetos and Andre Holzapfel. Automatic transcription of turkish microtonal music. *Journal of the Acoustical Society of America*, 138(3):accepted, September 2015. accepted.

Emmanouil Benetos and Tillman Weyde. Explicit duration hidden markov models for multiple-instrument polyphonic music transcription. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, November 2013. URL http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/117_Paper.pdf.

Emmanouil Benetos and Tillman Weyde. An efficient temporally-constrained probabilistic model for multiple-instrument music transcription. In M. Mueller and F. Wiering, editors, *16th International Society for Music Information Retrieval Conference*, pages 701–707, Malaga, Spain, October 2015. ISMIR.

Emmanouil Benetos, Sebastian Ewert, and Tillman Weyde. Automatic transcription of pitched and unpitched sounds from polyphonic music. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2014.

Taylor Berg-Kirkpatrick, Jacob Andreas, and Dan Klein. Unsupervised transcription of piano music. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1538–1546. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5432-unsupervised-transcription-of-piano-music.pdf.

Jeffrey Adam Bilmes. Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm. Master's thesis, Massachusetts Institute of Technology, September 1993.

Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma,

Justin Salamon, José Zapata, and Xavier Serra. ESSENTIA: an Audio Analysis Library for Music Information Retrieval. In *International Society for Music Information Retrieval Conference (ISMIR'13)*, pages 493–498, Curitiba, Brazil, November 2013.

Ciril Bohak and Matija Marolt. Finding Repeating Stanzas in Folk Songs. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 451–456, Porto, Portugal, 2012. ISBN 978-972-752-144-9.

Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Discriminative non-negative matrix factorization for multiple pitch estimation. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, Porto, Portugal, October 2012. URL http://ismir2012.ismir.net/event/papers/205-ismir-2012.pdf.

Albert S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, USA, 1990.

Kodály Center. The american folk song collection, 2004. URL kodaly.hnu.edu/home.cfm.

Chris Chafe, David A. Jaffe, Kyle Kashima, Bernard Mont-Reynaud, and Julius O. Smith. Techniques for note identification in polyphonic music. In *Proceddings of the 1985 International Computer Music Conference*, Burnaby, B.C., Canada, 1985. International Computer Music Association. URL https://ccrma.stanford.edu/files/papers/stanm29.pdf.

Wei Chai and Barry Vercoe. Music thumbnailing via structural analysis. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, MULTIMEDIA '03, pages 223–226, New York, NY, USA, 2003. ACM. ISBN 1-58113-722-2. doi: 10.1145/957013.957057. URL http://doi.acm.org/10.1145/957013.957057.

Tian Cheng, Simon Dixon, and Matthias Mauch. A deterministic annealing em algorithm for automatic music transcription. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, November 2013. URL http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/155_Paper.pdf.

Stephen Chu and Beth Logan. Music summarization using key phrases. Technical report, Cambridge Research Laboratory, April 2000.

Arshia Cont. Realtime multiple pitch observation using sparse non-negative constraints. In *Proceedings of the 7th International Conference on Music Information Retrieval*, Victoria (BC), Canada, October 2006. URL http://ismir2006.ismir.net/PAPERS/ISMIR06170_Paper.pdf.

Matthew Cooper and Jonathan Foote. Summarizing popular music via structural similarity analysis. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 127–30, New Paltz, NY, United States, 2003.

Matthew L. Cooper and Jonathan Foote. Automatic music summarization via similarity analysis. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 81–85, Paris, France, October 2002.

Manuel Davy and Simon J. Godsill. Bayesian harmonic models for musical signal analysis. In *Bayesian Statistics 7: the Seventh Valencia International Meeting*, Tenerife, Spain, 2003.

Iréne Deliége, Marc Mèlen, and Ian Cross. Musical schemata in real-time listening to a piece of music. *Music Perception*, 14(2):117–160, 1996.

Arnaud Dessein, Arshia Cont, and Guillaume Lemaitre. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 489–494, Utrecht, The Netherlands, August 2010. URL http://ismir2010.ismir.net/proceedings/ismir2010-83.pdf.

Diana Deutsch, editor. *The psychology of music*. Academic Press, New York, 1982.

John Stephen Downie. The Music Information Retrieval Evaluation eXchange (2005–2007): A Window Into Music Information Retrieval Research. *Acoustical Science and Technology*, 29(4):247–255, 2008. doi: 10.1250/ast.29.247.

J. P. Eckmann, S. O. Kamphorst, and D. Ruelle. Recurrence plots of dynamical systems. *EPL (Europhys. Lett.)*, 4: 973–977, 1987.

Antti Eronen. Chorus detection with combined use of mfcc and chroma features and image processing filters. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 229–36, Bordeaux, France, 2007.

Leonhard Euler. *Tentamen novae theoriae musicae ex certissismis harmoniae principiis dilucide expositae*. Saint Petersberg Academy, 1739.

Sam Ferguson and Densil Cabrera. Auditory spectral summarisation for audio signals with musical applications. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 567–572, Kobe, Japan, October 2009. URL http://ismir2009.ismir.net/proceedings/OS7-5.pdf.

Derry Fitzgerald. *Automatic Drum Transcription and Source Separation*. PhD thesis, Dublin Institute of Technology, 2004.

Jonathan Foote. Visualizing Music and Audio Using Self-Similarity. *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1) - MULTIMEDIA '99*, page 77–80, 1999. doi: 10.1145/319463.319472. URL http://portal.acm.org/citation.cfm?doid=319463.319472.

Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*. IEEE, 2000. doi: 10.1109/icme.2000.869637. URL http://dx.doi.org/10.1109/ICME.2000.869637.

Jonathan T. Foote and Matthew L. Cooper. Media segmentation using self-similarity decomposition. In *Proceedings of SPIE Storage and Retrieval for Multimedia Databases*, pages 167–175, 2003.

Jonathan T. Foote, Matthew L. Cooper, and Unjung Nam. Audio retrieval by rhythmic similarity. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, Paris, France, October 2002.

Benoit Fuentes, Roland Badeau, and Gaël Richard. Harmonic adaptive latent component analysis of audio and application to music transcription. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21 (9):1854–1866, September 2013. ISSN 1558-7916. doi: 10.1109/TASL.2013.2260741.

Simon Godsill and Manuel Davy. Bayesian harmonic models for musical pitch estimation and analysis. In *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, volume 2, pages 1769–1772, 2002.

Darryl Godsmark and Guy J. Brown. A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27(3–4):351–366, 1999. ISSN 0167-6393. doi: http://dx.doi.org/10.1016/S0167-6393(98)00082-X.

Masataka Goto. A predominant-fo estimation method for real-world musical audio signals: Map estimation for incorporating prior knowledge about fos and tone models. In *Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis*, Aalborg, Denmark, 2001.

Masataka Goto. A chorus-section detecting method for musical audio signals. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, page 50, October 2003.

Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1783–94, 2006.

Masataka. Goto and Yoichi Muraoka. A sound source separation system for percussion instruments. In *Transactions of the Institute of Electronics, Information and Communication Engineers*, volume J77-D-II (5), pages 901–911, 1994.

Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical and jazz music databases. In *ISMIR 2002, 3rd International Conference on Music Information Retrieval, Paris, France, October 13-17, 2002, Proceedings*, 2002. URL http://ismir2002.ismir.net/proceedings/03-SP04-1.pdf.

Graham Grindlay and Daniel P.W. Ellis. A probabilistic subspace model for multi-instrument polyphonic transcription. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 21–26, Utrecht, The Netherlands, August 2010. URL http://ismir2010.ismir.net/proceedings/ismir2010-5.pdf.

Peter Grosche, Björn Schuller, Meinard Müller, and Gerhard Rigoll. Automatic transcription of recorded music. *Acta Acustica united with Acustica*, 98(2):199–215, 2012.

Emilia Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra, 2006.

Kristoffer Jensen. A causal rhythm grouping. In *Computer Music Modeling and Retrieval*, Lecture Notes in Computer Science, pages 83–95. Springer Berlin / Heidelberg, 2005.

Kristoffer Jensen. Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal on Applied Signal Processing*, 2007(1):159–159, 2007. ISSN 1110-8657. doi: http://dx.doi.org/10.1155/2007/73205.

Florian Kaiser and Thomas Sikora. Music structure discovery in popular music using non-negative matrix factorization. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)*, Utrecht, Netherlands, August 2010. Society for Music Information Retrieval.

Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama. Separation of harmonic structures based on tied gaussian mixture model and information criterion for concurrent sounds. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004*, pages 297–300, Montreal, Quebec, Canada, May 2004. doi: 10.1109/ICASSP.2004.1326822. URL http://dx.doi.org/10.1109/ICASSP.2004.1326822.

K. Kashino and H. Tanaka. A sound source separation system with the ability of automatic tone modeling. In *International Computer Music Conference*, pages 248–255, Tokyo, Japan, 1993.

Kunio Kashino, Kazuhiro Nakadai, Tomoyoshi Kinoshita, and Hidehiko Tanaka. Organization of hierarchical perceptual sounds: Music scene analysis with autonomous processing modules and a quantitative information integration mechanism. In *International Joint Conferences on Artificial Intelligence*, pages 158–164. Morgan Kaufmann, 1995.

Anssi Klapuri. Introduction to music transcription. In Anssi Klapuri and Manuel Davy, editors, *Signal Processing Methods for Music Transcription*, pages 3–20. Springer US, 2006. ISBN 978-0-387-30667-4. doi: 10.1007/0-387-32845-9_1. URL http://dx.doi.org/10.1007/0-387-32845-9_1.

Anssi Klapuri and Manuel Davy, editors. *Signal Processing Methods for Music Transcription*. Springer US, 2006.

Anssi P. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6), November 2003.

Anssi P. Klapuri. A perceptually motivated multiple-fo estimation method. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pages 291–294, October 2005. doi: 10.1109/ASPAA.2005.1540227.

Peter van Kranenburg, Jörg Garbers, Anja Volk, Frans Wiering, Louis Grijp, and Remeco C. Veltkamp. Towards integration of mir and folk song research. In *Proceedings of the 8th International Conference on Music Information Retrieval*, 2007. URL http://depot.knaw.nl/8413/1/pvankranenburg_paper_ismir2007.pdf.

Carol L. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford Psychology Series. Oxford University Press, USA, 1990. ISBN 9780198022152.

C. Lee. The perception of metrical structure: experimental evidence and a model. In P. Howell, R. West, and I. Cross, editors, *Representing Musical Structure*, pages 59–127. Academic Press, London, 1991.

P. Lepain. Polyphonic pitch extraction from musical signals. *Journal of New Music Research*, 28(4):296–309, 1999.

Fred Lerdahl and Ray Jackendoff. An overview of hierarchical structure in music. *Music Perception: An Interdisciplinary Journal*, 1(2):229–252, 1983.

Beth Logan and Stephen Chu. Music summarization using key phrases. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, volume 2, pages II749–II752 vol.2, 2000. doi: 10.1109/ICASSP.2000.859068.

Lie Lu, Muyuan Wang, and Hong-Jiang Zhang. Repeating pattern discovery and structure analysis from acoustic music data. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR)*, pages 275–82, New York, NY, United States, 2004.

Robert Crawford Maher. *An Approach for the Separation of Voices in Composite Musical Signals*. PhD thesis, University of Illinois, 1989.

Robert Crawford Maher. Evaluation of a method for separating digitized duet signals. *Journal of Audio Engineering Society*, 38(12):956–979, 1990. URL http://www.aes.org/e-lib/browse.cfm?elib=6001.

Matija Marolt. A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, 6(3):439–449, August 2004. URL http://doi.ieeecomputersociety.org/10.1109/TMM.2004.827507.

Matija Marolt. Probabilistic segmentation and labeling of ethnomusicological field recordings. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 75–80, Kobe, Japan, October 2009.

Matija Marolt. Automatic transcription of bell chiming recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):844–853, March 2012. ISSN 1558-7916. doi: 10.1109/TASL.2011.2166957.

Matija Marolt and Marieke Lefeber. It' s time for a song – transcribing recordings of bell-playing clocks. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 333–338, Utrecht, The Netherlands, August 2010. URL http://ismir2010.ismir.net/proceedings/ismir2010-57.pdf.

Keith D. Martin. Automatic transcription of simple polyphonic music: Robust front end processing. Technical report, MIT Media Laboratory Perceptual Computing Section, 1996.

Matthias Mauch and Sebastian Ewert. The audio degradation toolbox and its application to robustness evaluation. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, 2013.

Matthias Mauch, Katy C. Noland, and Simon Dixon. Using Musical Structure to Enhance Automatic Chord Transcription. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)*, pages 231–236, 2009.

Brian McFee and Daniel P. W. Ellis. Analyzing Song Structure With Spectral Clustering. In *Proceedings of 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, page 405–410, 2014a. URL http://www.terasoft.com.tw/conf/ismir2014/proceedings/T073_319_Paper.pdf.

Brian McFee and Daniel P. W. Ellis. Learning to Segment Songs With Ordinal Linear Discriminant Analysis. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, page 5197–5201, 2014b. ISSN 15206149. doi: 10.1109/ICASSP.2014.6854594.

Brian McFee, Oriol Nieto, and Juan Pablo Bello. Hierarhical evaluation of segment boundary detection. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 406–412, Malaga, Spain, October 2015.

David K. Mellinger. *Event Formation and Separation in Musical Sound*. PhD thesis, Stanford University, Stanford, CA, December 1991. URL https://ccrma.stanford.edu/files/papers/stanm77.pdf.

Paul Mermelstein. Distance measures for speech recognition: Psychological and instrumental. In *Pattern Recognition and Artificial Intelligence*, pages 374–388. Academic Press, New York, 1976.

Marvin Minsky. Music, mind, and meaning. *Computer Music Journal*, 5(3), 1981.

Carl Mitchell, Mary Harper, and Leah Jamieson. On the complexity of explicit duration hmm's. *IEEE Transactions on Speech and Audio Processing*, 3(3):213–217, May 1995. ISSN 1063-6676. doi: 10.1109/89.388149.

James A. Moorer. *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*. PhD thesis, Stanford University, Stanford, CA, 1975. URL https://ccrma.stanford.edu/files/papers/stanm3.pdf.

James A. Moorer. On the transcription of musical sound by computer. *Computer Music Journal*, 1(4):32–38, 1977.

Meinard Müller, Peter Grosche, and Frans Wiering. Robust Segmentation and Annotation of Folk Song Recordings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 735–740, Kobe, Japan, October 2009.

Meinard Müller, Peter Grosche, and Frans Wiering. Automated analysis of performance variations in folk song recordings. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR)*, pages 247–256, Philadelphia, Pennsylvania, USA, 2010.

Meinard Müller, Nanzhu Jiang, and Peter Grosche. A Robust Fitness Measure for Capturing Repetitions in Music Recordings With Applications to Audio Thumbnailing. *IEEE Transactions on Audio, Speech & Language Processing*, 21(3):531–543, 2013.

Meinard Müller. *Information Retrieval for Music and Motion*. Springer Berlin Heidelberg, 2007. doi: 10.1007/978-3-540-74048-3. URL http://dx.doi.org/10.1007/978-3-540-74048-3.

Meinard Müller and Peter Grosche. Automated Segmentation of Folk Song Field Recordings. In *Proceedings of the ITG Conference on Speech Communication*, page 26–29, Braunschweig, Germany, 2012.

Meinard Müller, Peter Grosche, and Nanzhu Jiang. A Segment-Based Fitness Measure for Capturing Repetitive Structures of Music Recordings. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, page 615–620, Miami, Florida, USA, 2011. URL http://www.mpi-inf.mpg.de/~pgrosche/publications/2011_MuellerGroscheJiang_AudioStructure_ISMIR.pdf.

Bernhard Niedermayer. Non-negative matrix division for the automatic transcription of polyphonic music. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 544–549, Philadelphia, USA, September 2008. URL http://ismir2008.ismir.net/papers/ISMIR2008_198.pdf.

Oriol Nieto and Juan Pablo Bello. Music Segment Similarity Using 2D-Fourier Magnitude Coefficients. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, page 664–668, 2014. ISSN 15206149. doi: 10.1109/ICASSP.2014.6853679.

Oriol Nieto and Tristan Jehan. Convex Non-Negative Matrix Factorization For Automatic Music Structure Identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 236–240, 2013. ISBN 9781479903566. URL https://files.nyu.edu/onc202/public/publications/Nieto-ICASSP13.pdf.

National Library of Australia. Australian folk songs, 1994. URL folkstream.com.

Yannis Panagakis, Constantine Kotropoulos, and Gonzalo R Arce. ℓ1 -Graph Based Music Structure Analysis. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, page 495–500, Miami, Florida, USA, 2011.

Jouni Paulus and Anssi Klapuri. Music structure analysis by finding repeated parts. In *In Proceedings of the 1st ACM Workshop on Audio and Music Computing for Multimedia (AMCMM 2006*, pages 59–68. ACM Press, 2006.

Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech & Language Processing*, 17(6):1159–70, 2009a. ISSN 1063-6676. doi: http://dx.doi.org/10.1109/TASL.2009.2020533.

Jouni Paulus and Anssi Klapuri. *Computer Music Modeling and Retrieval. Genesis of Meaning in Sound and Music: 5th International Symposium, CMMR 2008 Copenhagen, Denmark, May 19-23, 2008 Revised Papers*, chapter Labelling the Structural Parts of a Music Piece with Markov Models, pages 166–176. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009b. ISBN 978-3-642-02518-1. URL http://dx.doi.org/10.1007/978-3-642-02518-1_11.

Jouni Paulus and Tuomas Virtanen. Drum transcription with non-negative spectrogram factorisation. In *Proceedings of the 13th European Signal Processing Conference*, Antalya, Turkey, September 2005.

Jouni Paulus, Meinard Müller, and Anssi Klapuri. State of the art report: Audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 625–636, Utrecht, The Netherlands, August 2010. URL http://ismir2010.ismir.net/proceedings/ismir2010-107.pdf.

Uroš Paščinski. *Automatic transcription of polyphonic singing*. PhD thesis, University of Ljubljana, Faculty of Computer and Information Science, 2015.

Geoffroy Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 35–40, Vienna, Austria, 2007.

Geoffroy Peeters and Victor Bisot. Improving Music Structure Segmentation Using Lag-Priors. In *Proceedings of 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, page 337–342, 2014.

Geoffroy Peeters, Amaury La Burthe, and Xavier Rodet. Toward automatic music audio summary generation from signal analysis. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 94–100, Paris, France, October 2002.

Martin Piszczalski. *A Computational Model of Music Transcription*. PhD thesis, University of Michigan, 1986.

Lawrence Rabiner. A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77 (2):257–286, February 1989.

Lawrence Rabiner and Biing H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, united states ed edition, April 1993. ISBN 0130151572.

Stanisław A. Raczyński, Nobutaka Ono, and Shigeki Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 381–386, Vienna, Austria, September 2007. URL http://ismir2007.ismir.net/proceedings/ISMIR2007_p381_raczynski.pdf.

Matti P. Ryynänen and Anssi Klapuri. Polyphonic music transcription using note event modeling. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pages 319–322, October 2005. doi: 10.1109/ASPAA.2005.1540233.

Matti P. Ryynänen and Anssi P. Klapuri. Modelling of note events for singing transcription. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio*, page 6. MIT Press, 2004.

Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, Aug. 2012. doi: 10.1109/TASL.2012.2188515.

Helmut Schaffrath. The essen folksong collection. D. Huron (ed.), Stanford, CA, 1995. URL essen.themefinder.org.

W. Andrew Schloss. *On the Automatic Transcription of Percussive Music - From Acoustic Signal to High-Level Analysis*. PhD thesis, Stanford University, Stanford, CA, May 1985. URL https://ccrma.stanford.edu/files/papers/stanm27.pdf.

Björn Schuller, Florian Dibiasi, Florian Eyben, and Gerhard Rigoll. Music thumbnailing incorporating harmony- and rhythm structure. In Marcin Detyniecki, Ulrich Leiner, and Andreas Nürnberger, editors, *Adaptive Multimedia Retrieval. Identifying, Summarizing, and Recommending Image and Music*, volume 5811 of *Lecture Notes in Computer Science*, pages 78–88. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-14757-9. doi: 10.1007/978-3-642-14758-6_7. URL http://dx.doi.org/10.1007/978-3-642-14758-6_7.

Joan Serrà, Emilia Gómez, Perfecto Herrera, and Xavier Serra. Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification. *IEEE Transactions on Audio, Speech and Language Processing*, 16:1138–1151, August 2008. ISSN 1558-7916. doi: 10.1109/TASL.2008.924595.

Joan Serrà, Meinard Müller, Peter Grosche, and Josep Ll Arcos. Unsupervised Music Structure Annotation by Time Series Structure Features and Segment Similarity. *Multimedia, IEEE Transactions on*, 16 (5):1229–1240, August 2014. ISSN 1520-9210. doi: 10.1109/TMM.2014.2310701.

Joan Serrà, Meinard Müller, Peter Grosche, and Josep Ll. Arcos. Unsupervised Detection of Music Boundaries by Time Series Structure Features. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, page 1613–1619. AAAI Press, 2012. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/viewPaper/4907.

Paris Smaragdis. *Redundancy Reduction for Computational Audition, a Unifying Approach*. PhD thesis, Massachusetts Institute of Technology, 2001.

Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, pages 555–560, 2011. URL http://ismir2011.ismir.net/papers/PS4-14.pdf.

Andrew Sterian and Gregory H. Wakefield. A model-based approach to partial tracking for musical transcription. In *Proceedings of International Computer Music Conference*, Beijing, China, 1999.

Grega Strle and Matija Marolt. Conceptualizing the ethnomuse: Application of cidoc crm and frbr. In *Proceedings of CIDOC2007*, 2007.

Li Su and Yi-Hsuan Yang. Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription. In *International Symposium on Computer Music Multidisciplinary Research*, June 2015.

H. Takeda, T. Nishimoto, and S. Sagayama. Rhythm and tempo analysis toward automatic music transcription. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1317–IV–1320, April 2007. doi: 10.1109/ICASSP.2007.367320.

Andranick S. Tanguiane. *Artificial Perception and Music Recognition*. Lecture Notes in Artificial Intelligence. Springer-Verlag, 1993. ISBN 9780387573946.

Jean Taque-Sutcliffe, Stephen Downie, and Shane Dunne. Name that tune! an introduction to musical information retrieval. *Proceedings of the 21st Annual Conference*, pages 204 — 216, 1993.

Petri Toiviainen and Tuomas Eerola. Suomen kansan esävelmät. digital archive of finnish folk tunes. Finnish Literary Society, 2004. URL esavelmat.jyu.fi.

Tero Tolonen and Matti Karjalainen. A computationally efficient multipitch analysis model. *Speech and Audio Processing, IEEE Transactions on*, 8(6):708–716, November 2000. ISSN 1063-6676. doi: 10.1109/89.876309.

Ivan Turk and Boris Kavur. *Mousterian "bone flute" and other finds from Divje Babe I cave site in Slovenia*, chapter Paleolithic bone flutes - comparable material, pages 179 – 184. Reseach Centre of the Slovenian Academy of Sciences and Arts, Institute of Archaeology: Založba ZRC, Ljubljana, Slovenia, 1997.

Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary Detection in Music Structure Analysis using Convolutional Neural Networks. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 417 – 422, Taipei, Taiwan, 2014.

P. van Kranenburg, M.J. de Bruin, L.P. Grijp, and F. Wiering. The meertens tune collections. *Meertens Online Reports*, 2014. ISSN 2352-2135.

Peter van Kranenburg and George Tzanetakis. A computational approach to the modeling and employment of cognitive units of folk song melodies using audio recordings. In *Proceedings of the 11th International Conference on Music Perception and Cognition*, pages 794–797, 2010.

Ron J. Weiss and Juan Pablo Bello. Identifying repeated patterns in music using sparse convolutive non-negative matrix factorisation. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)*, Utrecht, Netherlands, August 2010.

Ron J. Weiss and Juan Pablo Bello. Unsupervised discovery of temporal structure in music. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1240–1251, October 2011. ISSN 1932-4553. doi: 10.1109/JSTSP.2011.2145356.

Changsheng Xu, Namunu C. Maddage, and Xi Shao. Automatic music classification and summarization. *Speech and Audio Processing, IEEE Transactions on*, 13(3):441–450, May 2005. ISSN 1063-6676. doi: 10.1109/TSA.2004.840939.

S. J. Young, N. H. Russell, and J. H. S. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. Technical report, Cambridge University Engineering Department, 1989.

Tong Zhang and Ramin Samadani. Automatic generation of music thumbnails. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 228–231, July 2007. doi: 10.1109/ICME.2007.4284628.